Improving comprehension of Knowledge Representation languages: a case study with Description Logics

Improving comprehension of Knowledge Representation languages: a case study with Description Logics

Paul Warren , Paul Mulholland , Trevor Collins , Enrico Motta

PII: S1071-5819(18)30506-8
DOI: https://doi.org/10.1016/j.ijhcs.2018.08.009
Reference: YIJHC 2240

To appear in: *International Journal of Human-Computer Studies*

Received date: 15 November 2017
Revised date: 29 June 2018
Accepted date: 31 August 2018

Please cite this article as: Paul Warren , Paul Mulholland , Trevor Collins , Enrico Motta , Improving comprehension of Knowledge Representation languages: a case study with Description Logics, *International Journal of Human-Computer Studies* (2018), doi: https://doi.org/10.1016/j.ijhcs.2018.08.009

Highlights

- Insights from psychology and the philosophy of language help understanding of how people comprehend and reason with Description Logics
- The use of natural language in knowledge representation languages can assist comprehension but also create ambiguity
- Alternative or additional Manchester OWL Syntax keywords can significantly improve comprehension
- An understanding of De Morgan's Laws and the analogous duality laws for restrictions would aid reasoning with Manchester OWL Syntax
- Future development of knowledge representation languages should take account of psychological theories of reasoning and of how natural language is used

Improving comprehension of Knowledge Representation languages: a case study with Description Logics

Paul Warren, Paul Mulholland, Trevor Collins, Enrico Motta
Knowledge Media Institute, The Open University, U.K.

Abstract

Knowledge representation languages are frequently difficult to understand, particularly for those not trained in formal logic. This is the case for Description Logics, which have been adopted for knowledge representation on the Web and in a number of application areas. This work looks at the difficulties experienced with Description Logics; and in particular with the widely-used Manchester OWL Syntax, which employs natural language keywords. The work comprises three studies. The first two identify a number of difficulties which users experience, e.g. with negated intersection, functional properties, the use of subproperties and restrictions. Insights from cognitive psychology and the study of language are applied to understand these difficulties. Whilst these difficulties are in part inherent in reasoning about logic, and Description Logics in particular, they are made worse by the syntax. In the third study, alternative syntactic constructs are proposed which demonstrate some improvement in accuracy and efficiency of comprehension. In addition to proposing alternative syntactic constructs, the work makes some suggestions regarding training and support systems for Description Logics.

Keywords: Description Logics; Manchester OWL Syntax; User studies; Psychological theories of reasoning

1 Introduction

Knowledge representation (KR) languages are in common use to describe domains ranging from biology to finance. These languages are typically used by both computer scientists and domain experts. Early KR languages frequently employed the frame-based paradigm, developed by Minsky (1975), and inspired to a certain extent by psychological considerations[1]. In Minsky's (1975) approach *frames*, i.e. individual entities, have *terminals*, occupied by *assignments*, either by default or explicitly. This is, to an extent, analogous to the model used in object-oriented programming. An initial proposal was that frame-based representation be used for the Semantic Web (Lassila & McGuinness, 2001).

In fact, an alternative paradigm was adopted for the Semantic Web, that of Description Logics (DLs). DLs, and a reason for their adoption, are explained in Section 2. Their adoption was in the form of a family of W3C-standardized languages known as OWL, a permuted acronym of Web Ontology Language (W3C, 2001). As a result, DLs are now the dominant languages for specifying ontologies (Warren et al., 2014b). Moreover, they have been heavily studied by logicians; their computational properties are well understood and efficient reasoners have been developed. The initial interest by logicians also meant that early syntaxes were influenced by formal logic, and were perceived as not being ideal for domain experts with little training in logic and computer science[2]. In response to this, the

[1] Although psychologically inspired, the frame-based approach has been placed on a rigorous logical foundation, e.g. see Kifer et al. (1995).
[2] As will be discussed later, recent work has challenged this view (Alharbi et al., 2017).

Manchester OWL Syntax (MOS) (Horridge et al., 2006) was developed, making use of English keywords[3]. MOS is now a widely-used syntax.

Throughout these developments, there has been relatively little research into the difficulties which users of knowledge representation languages experience and the benefits, or otherwise, of the use of natural language. The research described in this paper investigates the difficulties which people experience in understanding and reasoning with DLs, in particular when expressed using MOS. The work takes ideas from the theory of reasoning in cognitive psychology and from the study of language, and applies these ideas to understand the difficulties experienced.

Section 2 provides an introduction to DLs and to the difficulties experienced with their use. Section 3 then provides an overview of various theories of reasoning developed by psychologists, and also some ideas from the study of language. Section 4 discusses the methodology used in the subsequent studies. Section 5 then describes the initial, exploratory study. This study focussed on commonly used DL constructs and identified a number of difficulties. The second study, described in section 6, investigated these difficulties in more detail, and also looked at some additional DL constructs. These studies suggested some modifications to the syntax which were investigated in a third study, described in section 7. Finally, section 8 discusses some implications of the work and makes some proposals for future research.

2 Description Logics, OWL and the Manchester OWL Syntax

This section provides a brief overview of DLs and the MOS syntax. It also describes some previous research into the difficulties which ontology developers experience with DLs.

2.1 Description Logics – overview

DLs are based on subsets of First Order Logic (FOL). However, FOL is concerned with defining and manipulating *propositions*; quantifiers are used to define propositions and Boolean operators to combine and negate them. DLs provide a language for *individuals* and *classes*; restrictions are used to define classes and Boolean concept constructors to combine and complement those classes. This will be illustrated in the next subsection. For an introduction to the theory of DLs, including the OWL standard, see Baader et al. (2017).

A key feature of DLs is the use of the Open World Assumption (OWA). The absence of a fact from a DL knowledgebase does not imply that the fact is not true. This contrasts with the Closed World Assumption (CWA) commonly assumed in database usage. Thus, if *Jane Smith* is not specified as the employee of a particular company in a DL knowledgebase, we cannot assume that she is not a company employee. We can only know that *Jane Smith* is not an employee if this information is in the knowledgebase, or can be deduced from other knowledgebase information. Another aspect of the OWA is that two names may refer to the same entity; *Jane Smith* and *J. Smith* may be the same person. We can only know they are

[3] MOS can be regarded as a Controlled Natural Language (CNL). It is, however, very restricted in its use of English, and Kuhn (2014), in a survey of CNLs, regards MOS as not sufficiently natural to be classified as a CNL. A number of DL languages have been developed which might more genuinely be regarded as CNLs. Schwitter et al. (2008) provide a comparison of three of these. Warren (2017; Section 2.5.3) provides a discussion of CNLs for DLs.

different if the knowledgebase explicitly says so, or if this can be deduced from other information. The OWA was a major reason for the adoption of DLs as a knowledge representation language for use on the World Wide Web (WWW). Unlike knowledge in corporate databases, knowledge on the WWW is rarely complete. The OWA also makes DLs appropriate for certain application areas, e.g. biological research (Stevens et al., 2007). However, the OWA does present difficulties. Rector et al. (2004) claim that it is "the biggest single hurdle to understanding OWL and Description Logics".

DLs are concerned with three types of entities:

- classes
- individuals, or instances, which are members of the classes
- object properties, which are defined between members of particular classes

As an example, we might have classes *Person* and *Dog*, with individuals *Tom* and *Rover* respectively, and a property *has_pet*, so that we could include a fact in the knowledgebase: *Tom has_pet Rover*. Here *Tom* is the subject, and *Rover* the object of the property *has_pet*. The OWL standard also includes datatype properties, between individuals and literals. We do not discuss these further as we do not believe that the problems of comprehension[4] which they pose will be substantially different from those posed by object properties.

2.2 Syntax and standardization

Since DLs were designed by logicians, the initial syntaxes were based on formalisms from logic. An example of this is the 'German DL' syntax. This used ⊔, ⊓, and ¬ for union, intersection and complement. Krötzsch et al. (2012) note that, by analogy with logic, these three operations are also referred to as disjunction, conjunction and negation. The existential (∃) and universal quantifier (∀) symbols are used to represent restrictions. For example: $\exists P.X$ defines a class containing those individuals which are the subject of a property P possessing an object in X, i.e. all the individuals a for which an individual b exists, such that $a\ P\ b$ and $b \in X$. Note that, although we are using the symbol for existential quantification, we are dealing with classes, not propositions, and the symbol is being used in a different way to its use as a quantifier.

Similarly, $\forall P.X$ defines a class containing all elements a such that, either:

- whenever a is the subject of an instance of P, the object is in the class X, or
- a is never the subject of an instance of X.

The second possibility is known as the 'trivial satisfaction of the universal restriction'. It corresponds to the convention in logic that any property holds for every element of the empty set.

The adoption of DLs for use on the WWW led to the definition of the Web Ontology Language (OWL), and this was accompanied by a variety of alternative syntaxes to German DL. Some of these were perceived as being verbose, and MOS was developed to be both relatively succinct and intelligible to non-logicians (Horridge et al., 2006; Horridge & Patel-Schneider, 2008). The chief features of MOS, as used in the studies reported in this paper, are shown in Table 1. The key points to note are that:

[4] For brevity, we use the word comprehension to mean not just the interpretation of DL statements but also reasoning about those statements.

- *or*, *and* and *not* are used for union, intersection and complement. This is consistent with the use of the terms *disjunction*, *conjunction* and *negation*.
- *some* and *only* are used for the existential and universal restrictions. Examples of this usage are shown later.
- Properties can be defined to have the following characteristics: transitive; functional; inverse functional; symmetric; asymmetric; reflexive; and irreflexive. Only the first four of these are used in the work reported here.

MOS is used in the Protégé ontology editor[5], which has been widely adopted (Warren et al., 2014b).

Table 1 Subset of MOS used in the studies

	Syntax	**Semantics**
Entity declarations	*Class X*	*X* a class.
	Individual a	*a* an individual.
	Property P	*P* a property.
Class expressions	*X or Y*	union of *X* and *Y*.
	X and Y	intersection of *X* and *Y*.
	not X	complement of *X*.
Restrictions	*P some X*	the existential restriction, i.e. the class of individuals who are the subject of the property *P* with object in the class *X*.
	P only X	the universal restriction, i.e. the class of individuals which are the subject of the property *P*, with objects only in *X*, plus those individuals which are not the subject of *P*.
Class axioms	*X SubClassOf Y*	*X* a subclass of *Y*, i.e. if an individual is in *X*, it is also in *Y*.
	X EquivalentTo Y	*X* and *Y* are equivalent classes, i.e. if an individual is in *X*, it is also in *Y*, and vice-versa.
	X DisjointWith Y	*X* and *Y* are disjoint, i.e. if an individual is in *X* it is not in *Y*, and vice-versa.
	Z DisjointUnionOf W, X, Y …	*Z* comprises all the individuals in *W*, *X*, and *Y* …, and no other individuals. Moreover, *W*, *X*, and *Y* are pairwise-disjoint, i.e. there are no individuals contained in more than one of these classes.
	P Domain X	All subjects of the object property *P* are in class *X*. Equivalent to *P some Thing SubClassOf X*.
	P Range X	All objects of the object property *P* are in class *X*. Equivalent to *Thing SubClassOf P only X*.
Individual axioms	*a Type X*	*a* a member of class *X*.
	a DifferentFrom b	*a* and *b* different individuals.
Property axioms	*P SubPropertyOf Q*	*P* is a subproperty of *Q*, i.e. if *a P b*, then *a Q b*.
	P InverseOf Q	*P* and *Q* are mutually inverse properties.
	P Characteristics transitive …	P has property characteristics, e.g. transitive.

[5] http://protege.stanford.edu/

2.3 Difficulties using DLs

Ontology developers, particularly non-logicians, experience difficulties using DLs. Rector et al. (2004), based on their experience of teaching OWL DL, identified several such difficulties. They were using 'Manchester House Style', a precursor to MOS which used *and* and *or* for conjunction and disjunction, and *someValuesFrom* and *allValuesFrom* for the existential and universal restrictions. The difficulties they identified included: a tendency to assume that *allValuesFrom* implies *someValuesFrom*, i.e. a tendency to overlook the second of the two possibilities associated with the universal restriction; confusion between *and* and *or*; and confusion between *P someValuesFrom (not X)* and *not (P someValuesFrom X)*.

Some difficulties arise from the keywords used. Other difficulties are inherent in DL but can be exacerbated or mitigated by choice of keywords. In the former category, the confusion between *and* and *or* is caused by the particular choice of these keywords; as will be argued later, *intersection* and *union* are less ambiguous. In the latter category, the tendency to overlook the 'trivial satisfaction of the universal restriction', as discussed in subsection 2.2, is inherent in DL but, as also will be discussed later, can be made worse or mitigated by the choice of syntax.

Rector et al. (2004) also recommended writing paraphrases of OWL statements. In these paraphrases, in anticipation of the later MOS, the universal restriction was represented using *only*. The existential restriction was represented using a combination of keywords. For example, a paraphrase of 'Pizza restriction (hasTopping someValuesFrom Tomato)' would be '*any* pizza which, *amongst other things*, has *some* tomato topping'. The inclusion of the phrase *amongst other things* recognises a user difficulty which will be further discussed in Sections 6 and 7.

Difficulties of comprehension can arise when debugging ontologies. After executing a reasoner, ontology developers may be confronted with an unexpected inference, or *entailment*. Whilst such an inference will be a logical consequence of the ontology axioms, the developer may regard it as incorrect from the domain perspective. Typically, ontology development systems can then be requested to provide a *justification*, defined as "a minimal subset of an ontology that is sufficient for an entailment to hold" (Horridge et al., 2011). However, understanding why the justification leads to the entailment can be difficult. Horridge et al. (2011) developed an intuitive model for the cognitive complexity of a justification and compared this with the difficulty actually experienced. They did this by presenting study participants with some justifications and corresponding putative entailments, and asking participants to indicate whether the entailment was, or was not, valid. They used the German DL syntax, with abstract names, e.g. *C1*, *C2* for classes and *prop1*, *prop2* for properties, which avoided participants making use of pre-existing domain knowledge. Their model "fared reasonably well" in predicting which questions study participants would find difficult and which they would find easy.

Nguyen et al. (2012) were interested in providing proof trees, in English, to explain why an entailment follows from a justification. They identified 51 deduction steps which could be used to create the proof trees and tested out the comprehensibility of these deduction steps on study participants. To do this, they presented a set of axioms and a putative inference and asked participants to confirm or refute the conclusion. However, to prevent the influence of pre-existing domain knowledge, they used a combination of meaningless words ('kalamanthis', 'tendriculos'), meaningful words ('plant', 'animal') and also words which are

not real English but have a semblance of being real words ('merfolk', 'lizardfolk'). Performance on these deduction rules varied widely. The easiest achieved 100% correct responses; the most difficult 4%. The latter required an understanding of the trivial satisfaction of the universal restriction.

3 Human reasoning and human language

This section describes the insights from reasoning studies and language studies which are used to guide and explain the work to be described later.

3.1 Theories of reasoning

Early theories of reasoning assumed "an unconscious logical calculus" which later was assumed to contain "formal rules of inference" (Johnson-Laird, 2010; page 194). The expectation was that people reason using formal rules of logic, similar to those used by a trained logician. These theories are variously referred to as *sentential*, *rule-based* (Stenning & Yule, 1997) or *mental logic* (Oaksford & Chater, 2001). The phrase *rule-based* will be used here, to emphasize the difference from the *model-based* theory described later.

An example of a rule-based theory is that developed by Rips (1983) for propositional reasoning. The theory employs a set of logical rules. Associated with each rule is an 'availability parameter', representing the probability of being able to retrieve and use the rule. From these parameters, the probability of being able to construct a particular chain of reasoning can be calculated. In his study, Rips presented participants with questions consisting of axioms and a putative inference. Participants were required to indicate whether the inference was "necessarily true" or "not necessarily true". Based on the results of this experiment, Rips estimated the availability parameter for each of his rules.

Braine (1978) provides another example of the rule-based theory applied to propositional logic. His theory accepts that the rules of human reasoning may not always correspond to those used in standard logic. For example, he argues that *if p then q* has a directionality, from *p* to *q*, in natural language which is absent from standard logic. In everyday discourse we are not concerned with what happens when *p* is not true and would not normally associate a truth value with the statement when *p* is false. Braine's (1978) natural propositional logic is based on the rules he claims we ordinarily use.

However, people do not always reason entirely by the application of formal rules. A classic early example of this is Wason's selection task (Wason, 1968), where participants are required to interpret rules in order to correctly select cards. The model-based theory was developed to explain this and other experimental results. The essence of this theory is that people construct mental models of reality and then require any inferences to be consistent with those models. Ehrlich and Johnson-Laird (1982) describe an early attempt to use mental models to explain experimental results relating to the layout of objects in two-dimensional space. However, mental models can be used to represent more abstract situations. Bucciarelli and Johnson-Laird (1999), for example, interpret relative difficulties with syllogisms in terms of mental model theory. Johnson-Laird (2005) provides an overview of mental model theory, whilst Johnson-Laird (2004) puts the theory in its historical context, tracing it back to the work of the American logician C.S. Peirce, through Wittgenstein's

(1922) picture theory of meaning, and Craik's (1967) view that cognition is based on forming models of the world.

As an example, consider conjunction, e.g. *there is a circle and there is a triangle*. Using C and T to represent circle and triangle, this is represented by one mental model:

C T

Exclusive disjunction, *there is a circle or there is a triangle, but not both*, requires two mental models:

C

T

Inclusive disjunction, *there is a circle or there is a triangle, or both*, requires three mental models:

C T

C

T

When dealing with a statement in propositional logic, the set of mental models corresponds to an expression in disjunctive normal form, with each mental model corresponding to a disjunct (Johnson-Laird et al., 1992). The essence of the mental model theory is that, when there is a requirement for more than one mental model, human reasoners are apt to overlook one or more of the models. This leads to the kinds of errors which humans frequently display.

The distinction between the rule-based and mental model theories is analogous to that between syntactic and semantic approaches in logic. Indeed, Braine and O'Brien (1998) use the phrase "syntax of thought" when writing about their rule-based theory. In contrast, Johnson-Laird and Byrne (1991, Prologue) use the phrase "an internal representation".

In addition to the rule-based and model-based theories, Halford et al. (1998) have developed a theory of reasoning based on ascribing a *relational complex*ity (RC) to each reasoning step. They give the following example: *John is taller than Mary* and *Mary is taller than Sue*, leading to the inference *John is taller than Sue*. This involves maintaining three items in working memory at the same time, and hence the RC is three. Halford et al. (2005) found that there was no significant difference in accuracy or time between problems with RC two or three. However, problems of RC four were answered significantly less accurately and in significantly longer time than problems of RC three. Problems of RC five were not answered significantly better than chance. RC theory can be regarded as complementary to both the other theories.

3.2 The ambiguity of natural language

MOS makes use of *and* and *or* to represent intersection and union. This presumably arose because, if we take *P(x)* and *Q(x)* to be predicates representing membership of classes C_P and C_Q, then *(P and Q)(x)* represents membership of the intersection of those classes, whilst *(P or Q)(x)* represents membership of the union. However, as already noted, DL, unlike FOL, is concerned with classes. The following examples illustrate the ambiguity which arises when *and* and *or* are used to represent class operations. The examples are taken from Partee and Rooth (1983); the analysis is the authors'. First consider three sentences constructed using *and*:

1. Susan will retire and buy a farm.
2. John and Mary are in Chicago.
3. She was wearing a new and expensive dress.

(1) is clearly shorthand for the conjunction of two propositions:

(Susan will retire) and (Susan will buy a farm).

(2) can also be interpreted as a conjunction:

(John is in Chicago) and (Mary is in Chicago).

An alternative interpretation, which conveys the same meaning, is to regard *and* as representing union:

{John} ∪ {Mary} are in Chicago.

(3) can also be interpreted as a conjunction:

(She was wearing a new dress) and (she was wearing an expensive dress).

However, here *and* can be regarded as representing intersection, again without any change of meaning:

She was wearing a {new dress} ∩ {expensive dress}.

In each of these three examples, the underlying semantics of *and* is the conjunction of two propositions. However, in (2) and (3), an interpretation in terms of class operations may be more natural. Moreover, it could be argued that union, rather than intersection, is a more natural interpretation of *and*, since intersection is often achieved by simply juxtaposing adjectives, e.g.:

She was wearing a new, expensive dress.

(2) above used *and* to represent union. The use of *or* to represent union, as in MOS, would clearly be wrong:

2A John or Mary is in Chicago.

Here, *or* represents uncertainty, since 2A is shorthand for:

(John is in Chicago) or (Mary is in Chicago).

However, Partee and Roth (1983) give another example:

4. The department is looking for a phonologist or a phonetician.

The underlying semantics of *or* is again disjunction:

(The department is looking for a phonologist) or (the department is looking for a phonetician).

However, the sentence is ambiguous. Interpreting as exclusive *or* implies uncertainty, i.e. that one of the above statements in brackets is true, but not both. Interpreting as inclusive *or* allows the possibility that both statements can be true, and hence that the department may be looking for someone from either discipline:

The department is looking for {phonologist} ∪ {phonetician}.

In this last interpretation, *or* has the same meaning as in MOS.

Mendonça et al. (1998) provide evidence that *and* and *or* give rise to ambiguities in practice. An examination of the use of *and* in preferred terms and synonyms in the SNOMED medical terminology, as used at that time, revealed that for 50.7% of the cases, both subjects were

required; for 46.1%, one or both; and for 3.2%, one or the other. A similar analysis for *or* gave the breakdown: one or the other 50.2%; one or both 49.5%; both 0.3%.

3.3 The implicature

Grice (1975) has made the point that a speaker (or writer) may use language to convey ideas which are not logically implied by the words used but which nonetheless are understood by the listener (or reader). Grice coined the word *implicature* for such ideas. Braine (1978) observes that, in reasoning according to standard logic, it is necessary first to extract the "minimum commitments" from the premises, whereas "ordinary comprehension processes" make use of implicatures.

Implicatures give rise to problems in the comprehension of MOS. The most well-known of these is that *only* suggests the existence of at least one instance of the object, as discussed in section 2.3. In everyday language, the sentence 'John has only sons' is taken to imply that John does have sons. In MOS, the statement *has_child only male* includes the possibility of no children at all.

The effect of this implicature can be related to the mental model theory. *has_child only male* is represented by two mental models:

has_child male
has child ⊥

In the second of these ⊥ represents Nothing, i.e. equivalent to the empty set. It is this second model which is frequently overlooked, encouraged by the implicature associated with *only*.

There are two possible implicatures surrounding the word *some*, depending on context. The most commonly studied example is that *some* implies *some not*. The idea here is that, if I say 'some of the students are industrious' I wish to convey the idea that some are not; otherwise I would have said 'all of the students are industrious'. However, there is another implicature which is relevant when considering the use of *some* in MOS. If I am asked 'does John have any children', and I reply 'John has some sons', the listener would reasonably assume that he does not have any daughters. Similarly, the MOS statement *has_children some male* may give rise to the assumption that there can be no female children. As already noted in subsection 2.3, Rector et al. (2004) discussed the use of the phrase "amongst other things" in a paraphrase of the existential restriction, in order to mitigate this difficulty.

The effect of this implicature can also be related to the mental model theory. *has_children some male* is represented by two mental models:

has child male
has child male has_child ¬male

In the second of these, ¬ represents complement. The danger is that this second model, representing the possibility that there is also a female child may get ignored, and this danger is encouraged by the implicature just discussed.

A final point to note is that the difficulty is exacerbated by the OWA. If the CWA were adopted, the statement *has_children some male*, without a complementary *has_children some female*, would imply that there were only male children.

4 Methodology

The next three sections report three studies, all of which followed the same pattern. There was an initial section which requested information about the participants, e.g. their knowledge of logic. There were then three or four sections containing the questions. Finally, there was a section which enabled the participants to provide feedback. In the first study, the questions in each section were related to a particular ontology pattern; in the other two studies the questions in each section shared particular MOS features. In all three studies, each question consisted of a set of axioms and a putative conclusion. Participants were required to indicate whether the conclusion was valid or non-valid. At the beginning of each study, participants were given a handout which explained all the languages features required to answer the questions. They were able to read this handout before beginning the study and keep it for reference during the study. In study 3, where there were two variants of the questions, two different handouts were used. Participants were free to move through the studies at their own pace.

In the first two studies, the *SurveyExpression*[6] web tool was used to display the questions and record the responses. *Camtasia*[7] was run on the PC to record screen activity and the recordings were subsequently analysed to obtain timing data. In the third study, for the reason explained in subsection 4.3, this set-up was replaced with the *MediaLab* application from Empirosoft[8], which ran on the PC and recorded both the responses and the response times. Statistical analysis was conducted using the R statistical package (R Core Team, 2014).

The participants were computer scientists or domain experts familiar with ontologies, drawn from university and industrial research laboratories. More information about the background of the participants is provided in Warren et al. (2014; 2015; 2017).

The following subsections consider some specific methodological issues.

4.1 Avoiding the use of prior knowledge

It is generally assumed, in studies of this kind, that one should avoid any bias due to prior knowledge. Two approaches to this were discussed in subsection 2.3. Horridge et al. (2011) used abstract names, whilst Nguyen et al. (2012) used 'nonsense' statements combining made-up words and real words. A third approach is to use a real ontology, but one which the participants are unlikely to have met, perhaps created specifically for the study. An example is provided by Vigo et al. (2014), who use a potato ontology.

The first of these approaches does not reflect the way ontologists normally work and may introduce difficulties of memorisation which are not relevant to the difficulties of reasoning being studied. Moreover, the assumption that one should remove all bias due to prior

[6] www.surveyexpression.com
[7] https://www.techsmith.com/camtasia.html
[8] www.empirisoft.com

knowledge is questionable. For example, it is likely that the user of an ontology has an inherent understanding of the characteristics of an object property, e.g. whether it is transitive or symmetric. The *Characteristics* statement may serve to confirm any expectations, but is primarily for the benefit of the computer. Johnson-Laird et al. (1989, page 668) make the same point when they argue that the logical properties of a term, e.g. the transitivity and symmetry of "is in the same place as" emerges "without any need to use explicit statements of these properties in the form of meaning postulates". Similarly, users may also be naturally aware of the subsumption relations in their domain, or this may be emphasized by the naming convention, rather than needing to refer to *SubClassOf* statements.

The distinction needs to be made between the use of prior knowledge to arrive immediately at the correct response, without reasoning, which clearly needs to be avoided; and the use of prior knowledge which is normally available in the process of reasoning. Laboratory experiments should create the most ecologically valid environment, e.g. through object property names reflecting the property characteristics and class names reflecting the hierarchical structure.

Study 1 was based on published ontology patterns, and the names used were largely taken from those patterns. In studies 2 and 3, names were chosen to achieve ecological validity:

- In questions involving object properties with specific characteristics, the object property names are chosen to suggest those characteristics, e.g. *greater_than_or_equal_to* to suggest transitivity. These names need to be chosen with great care to avoid suggesting additional characteristics which are not required, e.g. the use of *has_sibling* would have suggested transitivity, but would also have suggested symmetry.
- In the questions where object properties were used to create restrictions, both the properties and the classes were given meaningful names.
- In questions relating to Boolean operators, for simplicity abstract names were used. However, the use of several characters in somes name was designed to reinforce any class subsumption relations, e.g. *X, X_1, X_1_A.* All class subsumptions were, of course, declared in the appropriate MOS statements.
- For simplicity, individuals were given only single character names, e.g. *a, b, c*. These names were used with the object properties *greater_than_or_equal_to* and *has_nearest_neighbour*, where the use of single names for individuals seemed consistent with practice in algebra and geometry.

Study participants were, in fact, told only to draw on the information provided, and not make use of any preconceptions associated with names. That this is not entirely possible was illustrated by a comment from a participant in study 2, who objected to the use of *has_nearest_neighbour* as a functional property, since a point may have more than one nearest neighbour.

That meaningful names may aid reasoning is illustrated by an example from study 3. In this study, a participant reported using the concept of 'grandchild', not mentioned in the question, but deduced from the nested use of *has_child.*

4.2 Statistical treatment

In line with the recommended practice of the American Psychological Association (2010), p values are normally reported to three decimal places. The standard convention is used that $p < 0.05$ is required for significance.

Significance, effect size and sample size are interrelated. Van Schaik and Weston (2016) note the need for researchers "to choose a smallest important effect size", i.e. the smallest effect size which is of interest. Their approach includes the possibility of a result being "unclear". Alternatively, it is possible, given a minimum effect size of interest, to determine what sample size would be required to achieve a given statistical power. Obtaining participants for studies such as these is difficult because of the requirement for prior background in computing or ontology development. This has limited sample size, and hence quite large effects were required for significance. This needs to be borne in mind when interpreting the results.

In each of the three studies, the response time data was found to be positively skewed, as is illustrated in Figure 1 for the first study. Statistical tests such as the t-test and ANOVA rely on the time distribution being approximately normal. In the case of non-normal data, one possibility is to use non-parametric statistical tests. However, Hopkins et al. (2009) note that a transformation to reduce skewness, followed by a parametric test provides greater statistical power at small sample size than does a non-parametric test. One approach is to use a transformation selected from Tukey's ladder of powers (Scott, 2012). Figure 2 shows that, for the study 1 response time data, choosing the log transformation from Tukey's ladder of powers reduces skewness and provides a distribution closer to normal.

From the standpoint of statistical comparisons between questions, what is important is not the distribution overall, but the distribution of response time data within the individual questions. It was found that, for the majority of questions from the three studies, the optimum choice from the ladder of powers was the log transformation, and for consistency the $\log_{10}$ transformation was used for all statistical tests where the response time data was required to be normal.

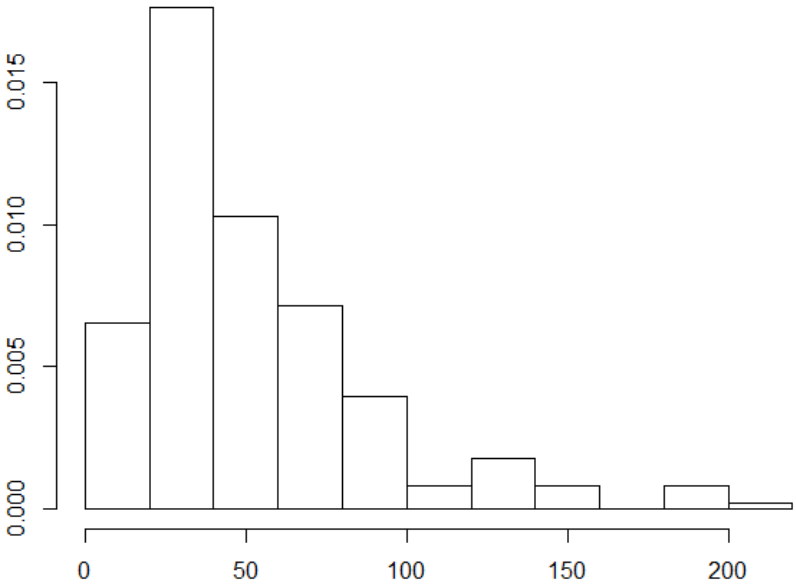

Figure 1 Probability density function for question response time (secs) – study 1

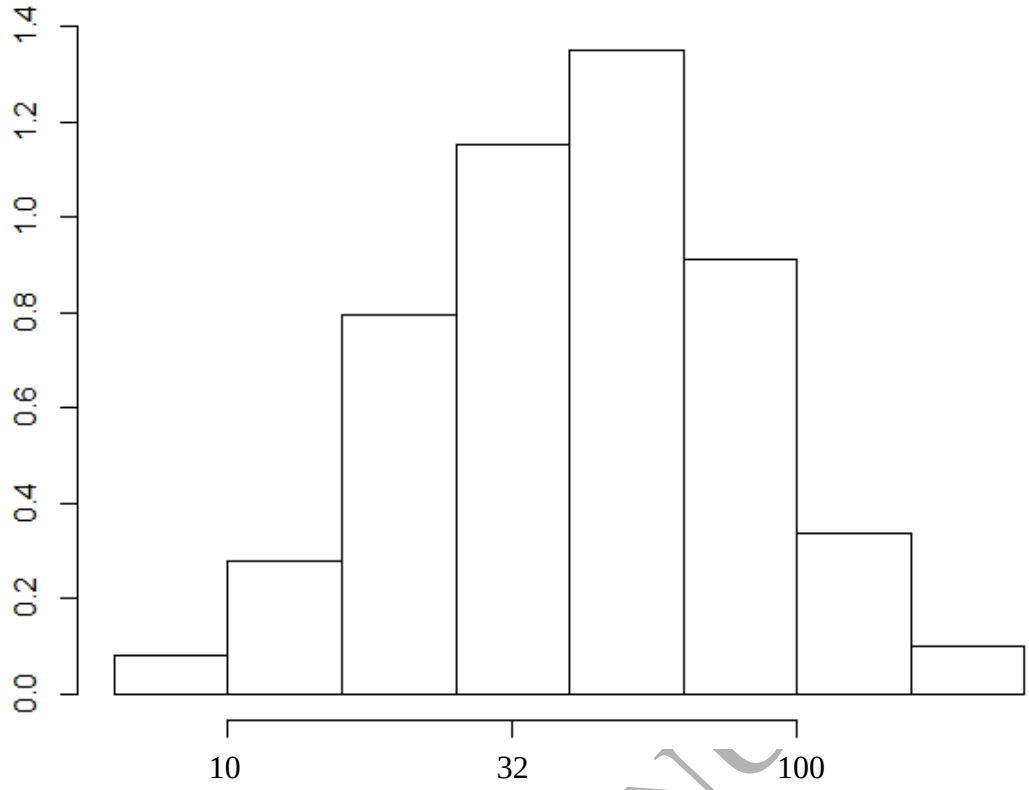

Figure 2 Probability density function for question response time after log transformation (secs) – study 1

4.3 Avoiding question order bias

The first, exploratory study, contained three sections, and hence six permutations of section order. There were twelve participants, and to compensate for any large-scale effect of question position, each section order was presented to two participants. The second study contained four sections, and hence 24 permutations of section order. There were 24 participants[9] and each section order was presented to one participant. In addition, in study 2 the questions in each section were presented in two reverse orders, with half the participants seeing the questions in one order and the other half seeing the alternative order. As explained in section 6, there appeared to be a considerable time penalty for the first question in each section, which was not sufficiently compensated for by the reversal of question order. This necessitated a careful choice of data for the statistical analysis. To avoid this problem, in the third study the use of the *MediaLab* application enabled the order of the sections and the order of the questions within the sections to be randomized.

5 Study 1 – investigating commonly used DL constructs

The object of this study was to explore participants ability to reason with commonly used DL constructs in the context of commonly used patterns. Warren et al. (2014) explain how the constructs and patterns were chosen, drawing on analyses by Power and Third (2010), Power (2010), Khan and Blomqvist (2010), and Warren et al. (2014). The patterns used were:

[9] This refers specifically to the participants from whom response time data was collected. As explained in section 6, there were four additional participants from whom only accuracy data was collected.

componency, *coparticipation*, and *types of entity*[10]. The latter two were extended so that all the required DL constructs were included in the study. Each pattern was the basis of a question section. For each section, participants were shown the pattern before moving on to the questions. The pattern was then repeated with each question, avoiding any need to memorise the pattern.

As the study was exploratory, unlike the two subsequent studies there were no specific hypotheses.

5.1 The *componency* pattern

Figure 3 shows the componency pattern, whilst Table 2 shows the related questions with accuracy and response time data.

Class Object	SubClassOf has_component only Object
	SubClassOf is_component_of only Object
Property has_part	Characteristics Transitive
Property is_part_of	Characteristics Transitive
	InverseOf has_part
Property has_component	SubPropertyOf has_part
Property is_component_of	SubPropertyOf is_part_of
	InverseOf has_component

Figure 3 Componency pattern

[10] Taken from the ODP portal: http://ontologydesignpatterns.org/wiki/Main_Page

Table 2 Componency pattern questions

No.	Question	validity	%age correct	mean time (s.d.) - secs
1*	A is_part_of B; C is_part_of B ⇒ A is_part_of C	not-valid	100%	62.3 (53.0)
2*	A is_part_of B; B is_part_of C ⇒ A is_part_of C	valid	100%	20.3 (9.7)
3*	B is_part_of C; A is_part_of B ⇒ A is_part_of C	valid	100%	30.7 (34.0)
4	A has_component B B has_component C ⇒ A has_component C	not-valid	33%	62.8 (36.2)
5*	A has_component B B has_component C ⇒ A has_part C	valid	83%	29.0 (15.8)
6*	A has_component B B is_part_of C ⇒ A has_part C	not valid	83%	57.9 (31.0)
7*	A has_component B C is_part_of B ⇒ A has_part C	valid	100%	37.4 (37.7)
8*	A Type Object A has_component B C Type not Object ⇒ B DifferentFrom C	valid	100%	49.9 (18.4)
9*	A Type Object; A has_part B C Type not Object ⇒ B DifferentFrom C	not valid	83%	47.5 (25.4)
10*	A has_component B C is_component_of B ⇒ C is_part_of A	valid	100%	54.2 (30.5)

** answered significantly better than chance ($p < 0.05$)*

With the exception of question 4, all questions were answered significantly better than chance. Question 4 was answered worse than chance, although not significantly so ($p = 0.194$, Fisher's Exact Test, one-sided). It appears that the participants who answered this question incorrectly have wrongly assumed that *has_component* is transitive. The error may have arisen because *has_component* is a subproperty of *has_part* which is transitive, and participants may have assumed that transitivity is inherited; despite the handout stating that transitivity is not inherited. In general, the inheritance of property characteristics cannot be assumed, although it is the case that some characteristics, e.g. functionality, are inherited. This issue is discussed further in Section 8, and in detail in Warren (2017; Chapter 2). Another factor may have been the everyday understanding of what it means to be a component, which conveys the implicature of transitivity. This illustrates the importance of choice of names, e.g. *has_direct_component* might avoid this implicature.

5.2 The extended *coparticipation* pattern

Figure 4 shows the extended coparticipation pattern, whilst Table 3 shows the related questions.

Class Event	EquivalentTo has_participant some Object
	DisjointWith Object
Class Object	*DisjointWith Event*
Class Player	*SubClassOf Object*
Class Game	*SubClassOf has_participant some Player*
Property coparticipates with	Domain Object, Range Object
	Characteristics Symmetric, Transitive
Property has_participant	Domain Event, Range Object
	InverseOf is_participant_in

N.B. the statements in italics were added to the original pattern.

Figure 4 Extended coparticipation pattern

Table 3 Extended coparticipation pattern questions

No.	Question	validity	%age correct	mean time (s.d.) - secs
1*	A coparticipates_with B ⇒ A Type not Event	valid	92%	54.9 (35.5)
2*	A is_participant_in B C coparticipates_with D ⇒ A DifferentFrom C	not valid	92%	68.8 (49.3)
3*	A is_participant_in B C is_participant_in B ⇒ A is participant in C	not valid	100%	43.6 (31.4)
4*	A has_participant B C is_participant_in D ⇒ B DifferentFrom D	valid	92%	44.6 (20.9)
5*	B coparticipates_with A B coparticipates_with C ⇒ C coparticipates with A	valid	100%	34.8 (15.9)
6	A Type Game ⇒ A Type Event	valid	67%	47.6 (25.6)

** answered significantly better than chance ($p < 0.05$)*

All questions were answered better than chance; all significantly so except question 6 ($p = 0.194$, Fisher's Exact Test, one-sided). It is not clear why this question was answered less well than the others. Possibly participants had difficulty with the existential restriction, although question 8 in the previous section involved the universal restriction and was answered well. Reasoning with restrictions was investigated more thoroughly in the second study.

5.3 The extended *types of entities* pattern

Figure 5 shows the extended types of entities pattern, whilst Table 4 shows the related questions.

Class Entity	EquivalentTo Event or Abstract or Quality or Object
Class Event	SubClassOf Entity
	DisjointWith Abstract, Quality, Object
Class Abstract	SubClassOf Entity
	DisjointWith Event, Quality, Object
Class Quality	SubClassOf Entity
	DisjointWith Event, Abstract, Object
Class Object	SubClassOf Entity
	DisjointWith Event, Abstract, Quality
Class Nonconceptual	*EquivalentTo Event or Object*
Class Nontemporal	*EquivalentTo Abstract or Quality or Object*
Property represents	*Characteristic Functional*

N.B. the statements in italics were added to the original pattern.

Figure 5 Extended types of entities pattern

Table 4 Extended types of entities pattern questions

No.	Question	validity	%age correct	mean time (s.d.) - secs
1*	A represents B; C represents D ⇒ A DifferentFrom C	not valid	83%	91.5 (61.7)
2	A Type Entity A Type not (Event and Quality) ⇒ A Type (Abstract or Object)	not valid	25%	75.1 (48.1)
3	A represents B; C represents D B Type Object; D Type Event ⇒ A DifferentFrom C	valid	50%	75.8 (31.0)
4*	A Type Entity A Type not (Event or Quality) ⇒ A Type Abstract or Object	valid	92%	44.0 (19.1)
5	A Type (Nonconceptual and Nontemporal) ⇒ A Type Object	valid	75%	63.1 (32.4)

** answered significantly better than chance ($p < 0.05$)*

Only two questions were answered significantly better than chance. Question 2 was answered worse than chance, although not significantly so ($p = 0.073$, Fisher's Exact Test, one-sided) and question 3 was answered at chance. These two questions are now considered in detail.

5.3.1 Negated intersection – question 2

Question 2 involves negated intersection. This question was answered significantly less accurately than question 4 which involved negated union ($p = 0.003$, Fisher's Exact Test, two-sided). Khemlani et al. (2012), working with naïve reasoners, found that 18% correctly answered questions relating to negated conjunction, whereas 89% correctly answered

questions relating to negated disjunction. These results are explicable in terms of mental model theory. Interpreting A and B as propositions or classes appropriately, then negated disjunction and negated union require one mental model:

¬A ¬B

Negated conjunction and negated intersection require three mental models:

¬A ¬B
¬A
¬B

It is possible that some people only create the first of these mental models, i.e. equivalent to the single mental model representing negated union.

5.3.2 Functional object property – question 3

Question 3 requires first an understanding that, since *B* and *D* are in disjoint classes (i.e. *Object* and *Event*), they must be different entities. It follows from the functionality of *represents*, that *A* and *C* must also be different entities. It may be that participants find it difficult to reason about functionality. However, this is a complex question; the final reasoning step is of RC 4, since it involves the entities *A*, *B*, *C*, *D*. It is not clear, therefore, to what extent the difficulty is due to an inherent difficulty of functionality and to what extent it is due to the complexity of the question.

5.4 Effect of participants' background knowledge

Participants were asked to rate their knowledge of formal logic, and of OWL or another DL formalism. In both cases they were provided with the scale: 'no knowledge'; 'a little knowledge'; 'some knowledge'; 'expert knowledge'. No participants rated themselves in the first category for either logic or DL. Table 5 shows the breakdown of responses between the other three categories, along with the percentage correct and the mean time to complete the study; the latter includes the time spent reading the on-screen preamble for the study overall and for each section.

Table 5 Participant performance by knowledge of formal logic, and by knowledge of OWL or another DL formalism

	Knowledge of logic				**Knowledge of OWL or other DL formalism**		
	no. participants	percentage correct	mean (s.d.) time - secs		no. participants	percentage correct	mean (s.d.) time - secs
A little knowledge	2	64%	1935 (968)		3	78%	2289 (384)
some knowledge	8	87%	1762 (796)		5	81%	1771 (922)
expert knowledge	2	88%	1107 (346)		4	90%	1113 (254)

As can be seen, there was increasing performance and decreasing time with increased knowledge both of formal logic and DL. There was a significant Spearman's rank correlation between knowledge of formal logic and accuracy of response ($\rho = 0.53$, $p = 0.038$, one-sided

test)[11] and between knowledge of DL and accuracy ($\rho = 0.54$, $p = 0.036$, one-sided test). There was also a significant Spearman's rank correlation between knowledge of DL and total time to answer the questions ($\rho = -0.65$, $p = 0.011$, one-sided test). However, the Spearman's rank correlation between knowledge of formal logic and total time was not significant ($\rho = -0.29$, $p = 0.178$). Here, the small number of participants in two of the categories makes reliable analysis difficult. It is also the case that there was a significant correlation for these participants between knowledge of formal logic and knowledge of DL ($\rho = 0.58$, $p = 0.024$, one-sided test) and this needs to be borne in mind when interpreting the data. More detail is given in Warren (2017, Chapter 6).

5.5 Conclusions from study 1

This first study identified four particular difficulties in reasoning with MOS:

1. an apparent assumption that transitivity is inherited by subproperties;
2. with existential restriction, in a situation which also involved the manipulation of subclasses;
3. in reasoning with negated intersection;
4. with a functional object property, in a situation of RC 4.

There was also evidence that increasing knowledge of formal logic and DL led to increasing accuracy and that knowledge of DL led to decreasing time to perform the study.

6 Study 2 – controlled comparisons

The second study built on the first, exploratory study, by using controlled comparisons between questions to further investigate participant difficulties. These comparisons enabled a better understanding, both of the effect of different DL constructs and also of increasing complexity.

6.1 Study 2 – practical details

The study followed the format described in Section 4. Response time data was collected from 24 participants, and accuracy data was collected from these 24 participants plus 4 additional participants. This arose because, on four occasions there were problems which led to a failure to collect all the response time data. Thus response time analysis was based on a sample of 24, accuracy analysis on a sample of 28.

There was no evidence that section position significantly affected accuracy ($\chi^2(3) = 3.1119$, $p = 0.375$)[12]. Moreover, there was no evidence that the position of a question within its section significantly affected accuracy ($\chi^2(9) = 11.8586$, $p = 0.221$)[12].

For response time, an ANOVA did reveal that there was a significant dependence of time on section position ($F(3,812) = 4.82$, $p = 0.002$). However, this should be compensated for by the use of every permutation of section position. A regression analysis of log response time against position of question within section also indicated a significant dependence ($F(1,814)$

[11] In this and subsequent usages of Spearman's rank correlation in this paper, the p-value could not be calculated exactly because of ties. It is assumed that this does not materially affect the results.

[12] This analysis was performed on data from 24 participants, representing each possible permutation of section order.

= 77.76, $p < 0.001$). This might be expected to be partially compensated by the use of the two reversed question orders, as discussed in section 4.4. However, inspection of the data revealed that there was an appreciable time premium for the first, and sometimes the second, question in each section. To avoid any bias this might introduce, where necessary analysis was conducted using subsets of the data.

6.2 Functional object properties

In study 1 participants had difficulty with a question involving a functional object property. This question section compares reasoning about functionality with reasoning about transitivity and investigates the effect of question complexity in reasoning about functionality. Only the former is reported here. The latter is reported in Warren et al. (2015).

6.2.1 Questions

Subsection 5.3.2 noted the difficulty which some participants had in the first study with a question involving a functional property. It was not clear to what extent the difficulty arose from the complexity of the question, and to what extent it was inherent in the concept of a functional property. To understand this better, this study undertook a comparison of questions involving functionality and transitivity, under conditions of controlled complexity.

The question in study 1 required a reasoning step of the form:

$$a\ F\ b;\ c\ F\ d;\ b\ \textit{DifferentFrom}\ d \Rightarrow a\ \textit{DifferentFrom}\ c \qquad (1)$$

Here, *F* is a functional object property and *a*, *b*, *c* and *d* are individuals. This step has RC 4 because it requires the concurrent attention to four individuals.

Another inference involving functionality is:

$$a\ F\ b;\ a\ F\ c \Rightarrow b\ \textit{SameAs}\ c \qquad (2)$$

This step has RC 3 because it only requires the concurrent attention to three individuals. If there is no inherent difference in difficulty between functionality and transitivity, then we might expect this reasoning step to display the same difficulty as the following one, where *T* represents a transitive object property:

$$a\ T\ b;\ b\ T\ c \Rightarrow a\ T\ c \qquad (3)$$

This also has RC 3 because it requires the concurrent attention to three individuals.

The two RC 3 inferences in (2) and (3) form the basis of a comparison between functionality and transitivity. Since these two inferences are individually relatively easy, it was thought that human reasoners might perform so well on them both as to make discrimination difficult. Therefore, an inference was used which involves functionality and requires two applications of the RC 3 reasoning step:

$$a\ F\ b;\ a\ F\ c;\ b\ F\ d;\ c\ F\ e \Rightarrow d\ \textit{SameAs}\ e \qquad (4)$$

This requires two reasoning steps of RC 3, linked by an additional reasoning step of RC 2 in which *c* is substituted for *b* in *b F d*, or *b* is substituted for *c* in *c F e*.

For comparison, an inference involving a transitive object property, *T*, also with reasoning steps of RC 3, 2 and 3, was constructed:

$$a\ T\ b;\ b\ T\ c;\ c\ SameAs\ d;\ d\ T\ e \Rightarrow a\ T\ e \qquad (5)$$

Here the axiom *c SameAs d* is used to necessitate a reasoning step of RC 2, analogous to the one which arises naturally in (4). Inference (5) is the basis of questions 1 and 2 in this section. Question 1 employs the correct, valid conclusion. Question 2 has a non-valid conclusion. Similarly, inference (4) is the basis of questions 3, with a valid conclusion, and question 4, with a non-valid conclusion. The remaining four questions in this section use a functional object property. Questions 5 and 6 replace the final reasoning step of (5) with a step of RC 4; questions 7 and 8 then replace the first reasoning step with one of RC 4. Note that correctly answering the questions containing functional properties requires an understanding of the OWA, specifically that different names may or may not refer to the same individual. Table 6 shows all eight questions, along with the accuracy and timing data. In the table, *T* and *F* are used to represent the object properties. In the actual questions *greater_than_or_equal_to* was used for transitivity and *has_nearest_neighbour* for functionality. For brevity this table, and all subsequent such tables, omit the declaration statements, e.g. property and individual declarations. These statements were included in the questions. As already explained, the questions were presented in two different orders. One group (1) of participants saw the questions in the order: 2, 3, 5, 8, 1, 4, 6, 7. The other group (2) saw the questions in the reverse order: 7, 6, 4, 1, 8, 5, 3, 2. The data for the two groups is also shown in the table.

Table 6 Questions employing transitivity and functional object properties

No.	**axioms**	**putative conclusion**	**valid / non-valid**	**RC**	**%age corr**			**mean time (s.d.) – secs**		
					overall N = 28	**order 1 N = 15**	**order 2 N = 13**	**overall N = 24**	**order 1 N = 12**	**order 2 N = 12**
1*	a T b; b T c; c SameAs d; d T e	a T e	valid	3,2,3	96%	100%	92%	34 (14)	28 (11)	40 (14)
2*		d T b	non-valid	n/a	86%	87%	85%	48 (34)	47 (16)	49 (47)
3*	a F b; a F c; b F d; c F e	d SameAs e	valid	3,2,3	75%	67%	85%	52 (36)	49 (33)	54 (39)
4*		a SameAs e	non-valid	n/a	96%	100%	92%	61 (46)	42 (26)	79 (56)
5	a F b; a F c; d F b; e F f; c DifferentFrom f	d DifferentFrom e	valid	3,2,4	61%	47%	77%	84 (67)	80 (79)	88 (55)
6*		a DifferentFrom d	non-valid	n/a	79%	73%	85%	92 (66)	73 (65)	111 (64)
7	a F b; c F d; b Differentfrom d; e F a; f F g; c SameAs g	e DifferentFrom f	valid	4,2,4	43%	40%	46%	109 (79)	86 (69)	132 (85)
8*		a DifferentFrom f	non-valid	n/a	71%	73%	69%	96 (47)	91 (47)	101 (48)

** answered significantly better than chance ($p < 0.05$); N = 28*

Questions 1 to 4 enabled the investigation of the hypothesis:

H2.1 Reasoning about functionality is inherently more difficult than reasoning about transitivity, after controlling for complexity.

6.2.2 Functionality versus transitivity – hypothesis H2.1

Comparison of questions 1 and 2 with questions 3 and 4 enables a comparison of the relative difficulty of transitivity and functionality. Considering first accuracy, Table 5 shows that, the transitive valid question was answered more accurately than the functional valid question, although not significantly ($p = 0.051$, Fisher's Exact Test, two-sided). Considering the non-valid questions, the situation was reversed with the functional question being answered more accurately than the transitive question, although again not significantly ($p = 0.352$, Fisher's Exact Test, two-sided).

Considering response time, the transitive valid question was answered significantly faster than the functional valid question ($t(23) = 3.0843$, $p = 0.005$, paired test, two-sided). The transitive non-valid question was answered faster than the functional non-valid question, but not significantly ($t(23) = 1.262$, $p = 0.220$, paired test, two-sided).

6.3 Complement, intersection and union

Study 1 showed that people have more difficulty with complemented intersection than with complemented union. The questions in this section were designed primarily to investigate the effect of alternative syntactic representations on reasoning with complemented intersection. This is discussed in detail in this section. Additionally, the questions looked at the effect of expanding complemented union, i.e. replacing *not (A or B)* with *not A and not B*; and with the effect of increasing complexity. These latter two effects are discussed in Warren et al. (2015).

6.3.1 Questions

In order to ground the questions in an ecologically valid context, they were based on Rector's (2003) use of Boolean operators to define exceptions in OWL. The questions are shown in Table 7. The table shows the minimum number of reasoning steps, and the maximum RC of these steps, for each of the valid questions. The table also shows the mental models associated with each set of axioms, giving both an initial form derived from the first axiom and then a form taking account of the subsequent axioms defining disjoint unions. In the representation of the mental models, *tc* is a representative member of *TOP_CLASS*, whilst *a*, *b*, *c*, *a1*, *a1x*, *a1y* are members of *A*, *B*, *C*, *A_1*, *A_1_X*, *A_1_Y* respectively. Questions 1 and 2, 3 to 8, and 9 and 10 have semantically equivalent axioms and the same mental models.

Table 7 Questions employing complement, union and intersection

no.	axioms	putative conclusion	valid / not-valid	no. steps (RC)	mental model(s)
1	Z EquivalentTo (TOP_CLASS and not A and not B) TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid	3 (3)	tc ¬a ¬b ≡ c
2	Z EquivalentTo (TOP_CLASS and not (A or B)) TOP_CLASS DisjointUnionOf A, B, C		valid	2 (3)	
3	Z EquivalentTo (TOP_CLASS and not (A and not A_1))	Z EquivalentTo (B or A_1)	valid	5 (3)	tc ¬a tc a1 ≡ b a1
4	TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A1, A2	Z EquivalentTo B	not-valid	n/a	
5	Z EquivalentTo (TOP_CLASS and (not A or A_1))	Z EquivalentTo (B or A_1)	valid	3 (2)	
6	TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo A_1	not-valid	n/a	
7	Z EquivalentTo ((TOP_CLASS and not A) or (TOP_CLASS and A_1))	Z EquivalentTo (B or A_1)	valid	5 (3)	
8	TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo A_2	not-valid	n/a	
9	Z EquivalentTo (TOP_CLASS and not (A and not (A_1 and not A_1_X)))	Z EquivalentTo (B or A_1_Y)	valid	9 (4)	tc ¬a tc a1 ¬a1x ≡ b a1y
10	TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2 A_1 DisjointUnionOf A_1_X, A_1_Y	Z EquivalentTo A_1_Y	not-valid	n/a	

One group (1) of participants saw the questions in the order: 3, 1, 8, 10, 5, 6, 9, 7, 2, 4. The other group (2) saw the questions in the reverse order: 4, 2, 7, 9, 6, 5, 10, 8, 1, 3. Table 8 shows the accuracy and response time data.

Table 8 Questions employing complement, union and intersection - accuracy and response time data

No.	valid / non-valid	%age correct			mean time (s.d.) – secs		
		overall N = 28	order 1 N = 15	order 2 N = 13	overall N = 24	order 1 N = 12	order 2 N = 12
1*	valid	82%	80%	85%	39 (26)	49 (31)	30 (15)
2*	valid	86%	100%	69%	43 (29)	22 (11)	64 (26)
3	valid	61%	60%	62%	96 (56)	128 (57)	63 (31)
4	non-valid	64%	53%	77%	105 (78)	42 (17)	167 (62)
5	valid	64%	67%	62%	65 (38)	60 (33)	69 (43)
6*	non-valid	79%	80%	77%	58 (33)	36 (16)	79 (32)
7*	valid	68%	80%	54%	70 (45)	39 (13)	101 (44)
8*	non-valid	89%	93%	85%	65 (26)	67 (29)	63 (24)
9	valid	54%	60%	46%	90 (48)	71 (31)	110 (56)
10*	non-valid	68%	60%	77%	94 (47)	94 (44)	95 (51)

** answered significantly better than chance ($p < 0.05$); N = 28*

Questions 3, 5 and 7 have semantically equivalent but syntactically different axioms and the same putative conclusion. This provides an opportunity to investigate the effect of syntax:

H2.2 Reasoning is influenced by syntax, in situations of semantic equivalence.

6.3.2 Effect of syntax – hypothesis H2.2

Inspection of questions 3, 5 and 7 shows a small amount of variation in accuracy. Question 3, the unexpanded form employing complemented intersection, was the least accurately answered. However, there was no significant difference between the questions ($\chi^2(2) = 0.31111$, $p = 0.856$).

Analysis of the response time data for questions 3, 5 and 7 needs to take account of the fact that question 3 was answered first by half the participants, and this appears to have attracted a considerable time penalty. Consequently, the statistical analysis was restricted to data for which questions 3, 5 and 7 occurred in the second half of the section (N = 12). On this basis, an ANOVA revealed no significant difference between the three questions ($F(2,33) = 2.0891$, $p = 0.140$).

A caveat is required here. This does not imply that syntactic form never makes a difference; simply that there was no significant difference for these particular questions. As will be shown in subsection 6.4.2, there are situations in which performance differs significantly between two semantically equivalent but syntactically different questions.

6.4 Complement and restrictions

This section is concerned with investigating the existential and universal restrictions, including interaction with complement. In particular, it enables an investigation of the effect of syntax and semantics on reasoning with these types of constructs.

6.4.1 Questions

Table 9 shows the questions, all of which have the same putative conclusion: *X DisjointTo Y*. Consider first questions 1 to 4. They share the same second axiom, which uses the universal restriction to constrain the class *Y*. For the first axiom, which constrains class *X*, questions 1 and 2 have the complement immediately before the named class (*MALE*), whilst questions 3 and 4 have the complement immediately before the anonymous class formed from the object property *has_child* and a restriction. Thus, all four possible variants of type of restriction (i.e. existential and universal) and position of complement are used. The first axioms of questions 1 and 4 are semantically equivalent, as are the first axioms of questions 2 and 3. To emphasise this equivalence, the first axioms of questions 1 and 4 are shown in distinctive typeface, and the first axioms of questions 2 and 3 in normal typeface. The semantic equivalence is based on the duality relations for restrictions. These are represented in MOS, where *R* is an object property and *C* is a class, as:

R some not C ≡ not R only C

R only not C ≡ not R some C

Questions 5 to 8 repeat the first axioms from the first four questions, and share a second axiom which uses the existential restriction to constrain the class Y.

Table 9 Questions employing existential and universal restrictions, and complement

No.	First axiom – constraining X	Second axiom – constraining Y	valid / non-valid
1	X SubClassOf has_child some (not MALE)	Y SubClassOf has_child only MALE	valid
2	X SubClassOf has_child only (not MALE)		non-valid
3	X SubClassOf not (has_child some MALE)		non-valid
4	X SubClassOf not (has_child only MALE)		valid
5	X SubClassOf has_child some (not MALE)	Y SubClassOf has_child some MALE	non-valid
6	X SubClassOf has_child only (not MALE)		valid
7	X SubClassOf not (has_child some MALE)		valid
8	X SubClassOf not (has_child only MALE)		non-valid

N.B. All questions have the same putative conclusion: X DisjointTo Y

One group (1) of participants saw the questions in the order: 1, 2, 3, 4, 5, 6, 7, 8. The other group (2) saw the questions in the reverse order: 8, 7, 6, 5, 4, 3, 2, 1. Table 10 shows the accuracy and response time data.

Table 10 Questions employing existential and universal restrictions, and complement - accuracy and response time data

No.	valid / non-valid	%age corr			mean time (s.d.) – secs		
		overall N = 28	order 1 N = 15	order 2 N = 13	overall N = 24	order 1 N = 12	order 2 N = 12
1	valid	61%	53%	69%	52 (39)	70 (44)	34 (25)
2	non-valid	50%	47%	54%	33 (18)	28 (12)	39 (22)
3*	non-valid	68%	67%	69%	45 (22)	43 (17)	47 (27)
4*	valid	75%	73%	77%	43 (25)	48 (28)	39 (22)
5	non-valid	64%	60%	69%	41 (30)	48 (36)	33 (22)
6	valid	50%	40%	62%	44 (40)	43 (49)	46 (30)
7*	valid	79%	80%	77%	43 (37)	32 (21)	54 (46)
8*	non-valid	68%	73%	62%	60 (37)	41 (28)	78 (37)

** answered significantly better than chance ($p < 0.05$); N = 28*

The fact that the first axioms fall into two groups of semantically equivalent axioms, combined with the two alternatives for the second axiom, means that the questions fall into four pairs of semantically equivalent questions: {1, 4}, {2, 3}, {5, 8}, {6, 7}.

6.4.2 Syntax versus semantics – H2.2

Inspection of Table 10 shows that, with the exceptions of questions 5 and 8, within each pair of semantically equivalent questions there is an appreciable difference in accuracy. Indeed, in the case of questions 6 and 7 the difference is significant ($p = 0.0496$, Fisher's Exact Test, two-sided). Thus it appears that syntax has here an appreciable effect on performance. A more detailed inspection of the data gives some insight into what determines performance. In questions 4 and 7, the questions which were answered most accurately, the first axiom uses the complement of the anonymous class used in the second axiom. Since complementary

classes are disjoint, *X* and *Y*, being subclasses of the two anonymous classes, will also be disjoint. Thus, it is immediately clear from the syntax that the conclusion is valid, without the need to construct detailed mental models.

In the other questions, the conclusion is not so obvious from the syntax, and more thought is required. In questions 2 there may also be confusion between *has_child only (not MALE)* and *not (has_child only MALE)*. If the latter were substituted for the former, then the participant would erroneously conclude that the conclusion is valid. A similar comment applies to question 5, where *has_child some (not MALE)* may be confused with *not (has_child some MALE)*. Indeed, Rector et al. (2004) have noted this source of confusion.

Question 2 also suffers from the difficulty that it requires an understanding of the trivial satisfaction of the universal restriction; another difficulty pointed out by Rector et al. (2004). In terms of mental model theory, this is equivalent to ignoring the second mental model in the representation of the universal restriction.

A final point to note is that, if participants had been aware of the duality relations shown in subsection 6.4.1, questions 1 and 6 would have been considerably easier, since they can be transformed into questions 4 and 7, the most accurately answered questions. Knowledge of these relations might also have guarded against the confusion of *only not* with *not ... only* and *some not* with *not ... some*, as may have occurred in questions 2 and 5.

6.5 Nested restrictions

This section is concerned with further investigation of the difficulties of reasoning with the existential and universal restrictions. An additional difficulty is introduced here by the use of nested restrictions, i.e. the object of one restriction is a class defined by a second restriction. The next subsection describes the questions, and the following two subsections discuss the results for the valid and non-valid questions.

6.5.1 Questions

Table 11 shows the questions. In the second column, there are axioms constraining the class *X*. For half the questions one axiom is used, containing two restrictions. For the other questions, two axioms are used connected by a named class, *Y*. As a consequence, the first axiom for question 1 is equivalent to the first two axioms for question 5. A similar relation holds between questions 2, 3, 4 and 6, 7, 8. Questions 1 to 4 use all the patterns *some ... some*, *some ... only*, *only ... some*, and *only ... only* and questions 5 to 8 repeat these patterns.

The third column then shows the remaining two axioms which are used to constrain an individual *a*. Questions 1 to 4 use *some not*, whilst questions 5 to 8 use *not ... some*. All the questions have the same putative conclusion: *a Type (not X)*.

Table 11 Questions with nested restrictions

No.	First axiom(s) – constraining X	Remaining axioms – constraining a	valid / non-valid
1	X SubClassOf (has_child some (has_child some FEMALE))	a has_child b; b Type has_child some (not FEMALE)	non-valid
2	X SubClassOf has_child some Y; Y EquivalentTo has_child only FEMALE		non-valid
3	X SubClassOf (has_child only (has_child some FEMALE))		non-valid
4	X SubClassOf has_child only Y; Y EquivalentTo has_child only FEMALE		valid
5	X SubClassOf has_child some Y; Y EquivalentTo has_child some FEMALE	a has_child b; b Type (not (has_child some FEMALE))	non-valid
6	X SubClassOf (has_child some (has_child only FEMALE)		non-valid
7	X SubClassOf has_child only Y; Y EquivalentTo has_child some FEMALE		valid
8	X SubClassOf (has_child only (has_child only FEMALE))		non-valid

N.B. All questions have the same putative conclusion: a Type (not X)

One group (1) of participants saw the questions in the order: 1, 2, 3, 4, 8, 7, 6, 5. The other group (2) saw the questions in the reverse order: 5, 6, 7, 8, 4, 3, 2, 1. Table 12 shows the accuracy and response time data.

Table 12 Questions with nested restrictions – accuracy and response time data

No.	valid / non-valid	%age corr			mean time (s.d.) – secs		
		overall N = 28	order 1 N = 15	order 2 N = 13	overall N = 24	order 1 N = 12	order 2 N = 12
1*	non-valid	71%	73%	69%	69 (45)	96 (45)	41 (23)
2	non-valid	57%	67%	46%	79 (53)	96 (56)	62 (46)
3*	non-valid	71%	73%	69%	63 (43)	71 (41)	56 (45)
4	valid	57%	67%	46%	63 (39)	55 (43)	71 (36)
5	non-valid	54%	47%	62%	88 (62)	47 (28)	128 (60)
6	non-valid	64%	73%	54%	73 (45)	51 (34)	95 (44)
7*	valid	71%	87%	54%	80 (36)	62 (36)	98 (26)
8	non-valid	50%	53%	46%	55 (30)	41 (22)	68 (32)

** answered significantly better than chance ($p < 0.05$); N = 28*

This section appears to be more difficult than the section with restrictions discussed in subsection 6.4. The overall accuracy for this section was 62%, and that for the previous section was 64%, although the difference was not significant ($p = 0.695$, Fisher's Exact Test, two-sided). However, the mean response time for this section was 71 seconds, compared with 45 seconds for the previous section, and this difference was significant ($F(1,382) = 47.868$, $p < 0.001$).

6.5.2 Questions with valid putative conclusions

There were two questions with valid putative conclusions: questions 4 and 7. Question 7 was one of the three questions answered significantly better than chance. It is clear from the syntax that the anonymous class defined in the final axiom, i.e. *not (has_child some FEMALE)*, is the complement of *Y*. This immediately implies that *b* cannot be in *Y*, and hence, given the use of the universal restriction in the first axiom, that *a* cannot be in *X*. This contrasts with question 4, which was answered less well. Here there is no such syntactic clue and participants are likely to find it more difficult to realise that *b* cannot be in *Y* and hence that *a* cannot be in *X*. Whilst for both questions the mean time to answer correctly was less than the mean time to answer incorrectly, the difference was only significant for question 7 (question 7: $t(16.932) = 3.11$, $p = 0.006$; question 4: $t(13.235) = 0.14744$, $p = 0.885$). This suggests that participants who answered question 7 correctly picked up the syntactic clue quickly. Note that if the final axiom of question 4 was transformed using the appropriate duality rule, to *b Type not* (*has_child only FEMALE)*, then this question would offer a syntactic clue, analogous to question 7.

6.5.3 Questions with non-valid putative conclusions

The remaining six questions all had non-valid conclusions. The best answered were questions 1 and 3. In both questions, the first axiom states that a member of the anonymous class formed with the second restriction has a female child and the final axiom states that *b* has a non-female child. There is no contradiction here, although it does require an awareness of the second mental model in the representation of the existential restriction. In both questions, having deduced that *b* can be in the anonymous class formed by the second restriction of the first axiom, it is clear that *a* can be in *X*.

Question 6 was also answered relatively well. The non-validity of the conclusion can be deduced in two ways. It is possible for *b* to be in the anonymous class formed with the second restriction in the first axiom, although this does require an understanding that the universal restriction can be trivially satisfied, i.e. it requires an understanding of the second model in the representation of the universal restriction. Although this was pointed out in the handout, participants might easily overlook this and revert to the more natural sense of *only*, with the implicature that a female child does exist. However, even if participants conclude wrongly that *b* cannot be in *Y*, it is still possible for *a* to be in *X* by having another child which does have a female child. This also requires an awareness of the second model in the expansion of the existential restriction. A participant who initially takes the first model for the first (existential) restriction, and then fails to take note of the second model for the second (universal) restriction, can still answer the question correctly by backtracking to the first restriction and taking the second model for this restriction.

In question 5 there is a strong syntactic clue that *b* cannot be in *Y*; *not (has_child some FEMALE)* is clearly the complement of *has_child some FEMALE*. To answer the question correctly, it is essential to backtrack to the first restriction and use its second model to understand that it is possible for *a* to be in *X* by having another child which itself does have a female child. A similar argument applies to question 2. It is relatively easy to see that *b* cannot be in *Y*, since *b* cannot have only female children and some non-female children. Again, to understand the non-validity of the conclusion it is essential to backtrack to the second model for the first restriction.

Finally, question 8 has similarities with question 6. It is possible for *b* to be in the anonymous class formed with the second restriction in the first axiom, and hence *a* in *X*, since there is no contradiction in logic between having only female children and having no children at all. However, unlike with question 6, if participants fail to understand this, backtracking to the first (universal) restriction offers no second chance. The second model in the interpretation of the first universal restriction is not applicable, since this would require that *a* has no children, and we know that *a has_child b*. So, although question 6 can be correctly answered by making use of the trivial satisfaction of the universal restriction, this knowledge is not essential. Only question 8 requires that knowledge for a correct answer.

6.6 Effect of participants' background knowledge

As with the previous study, participants were asked to rate their knowledge of formal logic and of OWL or another DL formalism. Table 13 shows the breakdown of participants, along with the mean accuracy and time in each category. A Spearman's rank correlation revealed no significant correlation between accuracy and knowledge of logic ($\rho = 0.19$, $p = 0.163$, one-sided), although there was a significant correlation between accuracy and knowledge of DL ($\rho = 0.41$, $p = 0.015$, one-sided). There was no significant correlation between time and knowledge of logic ($\rho = -0.13$, $p = 0.266$, one-sided) or knowledge of DL ($\rho = -0.29$, $p = 0.087$).

Table 13 Participant performance by knowledge of formal logic, and by knowledge of OWL or another DL formalism

	Knowledge of formal logic				**Knowledge of OWL or another DL formalism**			
	accuracy		**time**		**accuracy**		**time**	
	N	%age correct	N	mean time (s.d.) - secs	N	%age correct	N	mean time (s.d.) - secs
No knowledge	2	51%	1	2484 (NA)	3	72%	2	2545 (429)
A little knowledge	3	80%	3	2087 (217)	8	58%	7	2573 (871)
some knowledge	16	65%	14	2522 (816)	9	67%	7	2401 (776)
expert knowledge	7	77%	6	1929 (768)	8	80%	8	1965 (700)

6.7 Conclusions from study 2

Table 14 summarises the findings from study 2. The study identified that reasoning about functionality can take longer than reasoning about transitivity, after controlling for complexity. The study has further investigated difficulties with the universal and existential restrictions. In some cases, the difficulties appeared to stem from a failure to take account of both the mental models for these restrictions. This has been well known for the universal restriction, but also appears true for the existential restriction.

Table 14 Summary of study 2 findings

Hypothesis	Accuracy	Response time
H2.1 Functionality harder than transitivity	No significant effect.	Valid transitive question answered significantly faster than valid functional question.
H2.2 Reasoning performance is influenced by syntax	No significant effect for Boolean constructors. Significant difference between semantically equivalent but syntactically different forms involving restrictions.	No significant effect for Boolean constructors. For questions involving restrictions, no clear conclusions can be drawn.

Two general comments are appropriate. Firstly, reasoning with DL statements is very difficult. 15 of the 34 questions were not answered significantly better than chance. Secondly, participants appeared to use both semantic and syntactic approaches to reasoning.

Finally, prior knowledge had less effect on performance than was the case in the first study. Overall, the questions in the second study were harder than those in the first study, and this may be the reason for the reduced effect.

7 Study 3 – modifying the syntax

The previous two studies identified some of the difficulties which people experience when reasoning with DLs, in particular using MOS. This section looks at how extensions to the MOS vocabulary can mitigate those difficulties. These effects were evaluated by using questions isomorphic to some of those from the previous two studies. This enabled comparisons with the previous questions, chiefly with those in study 2.

In this study, as explained in Section 4, the *MediaLab* software was used, on the experimenter's computer, to pose the questions and to record the responses and response times. This enabled the section and question order to be randomized, to avoid any bias arising from question order. There were 30 participants and the study existed in two variants. Some, but not all, of the questions differed in the two variants. Participants were allocated alternately between the two variants.

Subsections 7.1 to 7.4 describe the four parts of the study. These are parallel to those in study 2, being concerned with: functional and inverse functional object properties; complement, union and intersection; complement and restrictions; and nested restrictions. Finally, subsection 7.5 draws some conclusions.

7.1 Functional and inverse functional object properties

Study 2 demonstrated that valid questions employing functional object properties took significantly longer than valid questions employing transitive object properties. At least one participant in a previous study had appeared to be confused between functionality and inverse

functionality, i.e. to be unclear whether it was the subject or object of the property which was unique. It was hypothesized that the introduction of an additional keyword, *solely*, before the object would emphasize that it is the object which is unique and facilitate reasoning. This leads to the hypothesis:

> H3.1 The introduction of the additional keyword *solely* between a functional object property and its object will improve participant performance.

Neither of the previous studies employed inverse functional object properties. It was thought that reasoning with such properties was likely to display the same kind of difficulties as reasoning with functional object properties; and hypothesized that the use of the keyword *solely*, this time before the subject, would aid reasoning. This led to the hypothesis:

> H3.2 The introduction of the additional keyword *solely* before the subject of an inverse functional property will improve participant performance.

Two questions were designed using inverse functional properties. As there were no questions from previous studies with which to make a comparison, half the participants saw these questions in standard MOS, the other half saw the questions with the inclusion of *solely*.

7.1.1 Functional object properties – hypothesis H3.1

Table 15 shows the six study 3 questions using functional object properties. For these questions, there was no difference between the two variants of the study, i.e. all participants saw the same questions. In the table, *F* is used for brevity to represent the object property. As with study 2, in the actual questions, *has_nearest_neighbour* was used. These questions are isomorphic to questions 3 to 8 in Table 6, differing only in the use of different individual names and the use of the keyword *solely*.

Table 15 Questions employing functional object properties

	axioms (F = *has_nearest_neighbour*)	**putative conclusion**	**validity**	**relational complexity**
1	r F solely s; r F solely t;	v sameAs w	valid	3,2,3
2	s F solely v; t F solely w	r sameAs t	not valid	n/a
3	r F solely s; r F solely t; v F solely s;	v DifferentFrom w	valid	3,2,4
4	w F solely x; t DifferentFrom x	r DifferentFrom v	not valid	n/a
5	r F solely s; t F solely v;	w DifferentFrom x	valid	4,2,4
6	s DifferentFrom v; w F solely r; x F solely z; t SameAs z	r DifferentFrom x	not valid	n/a

Table 16 shows the accuracy and response time data for the six questions. In each case, the comparable data from study 2 is also shown. For the six questions aggregated, the accuracy in study 3 was 76%, compared with 71% in study 2; this difference was not significant ($p = 0.334$, Fisher's Exact Test, two-sided). Comparison of the two studies shows that for each of the valid questions the accuracy was greater in study 3 than in study 2. This suggests comparing the three valid questions aggregated, for which the accuracy in study 3 was 72% compared with 60% in study 2. However, again the difference was not significant ($p = 0.081$, Fisher's Exact Test, two-sided).

Table 16 Questions employing functional object properties – accuracy and response time data*

	study 2 – without *solely* **from Table 6**		**study 3** – with *solely*	
	%age correct N = 28	mean time (SD) – secs N = 24	%age correct N = 30	mean time (SD) – secs N = 30
1	75%	52 (36)	83%	39 (31)
2	96%	61 (46)	83%	50 (29)
3	61%	84 (67)	70%	58 (27)
4	79%	92 (66)	83%	78 (49)
5	43%	109 (79)	63%	73 (37)
6	71%	96 (47)	70%	90 (46)
All questions	71%	83 (61)	76%	65 (41)
valid questions	60%	81 (67)	72%	57 (35)
non-valid questions	82%	83 (55)	79%	73 (45)

** These questions are numbered as in study 3; the analogous questions in study 2 were numbered from 3 to 8.*

Comparison of the response times in Table 16 shows that, for each question, the mean response time for study 3 was less than for study 2. For each pair of questions of equivalent complexity, this difference was greater for the valid question. When the valid questions were aggregated, the difference between the two studies was significant ($t(144.7) = 2.8373$, $p = 0.005$). For the non-valid questions, the difference was not significant ($t(148.61) = 1.0043$, $p = 0.317$)[13].

7.1.2 Inverse functional object properties – hypothesis H3.2

Table 17 shows questions 7 and 8, which were used to investigate the effect of *solely* in the case of inverse functional object properties. Variant 1 uses conventional MOS. Variant 2 is identical except for the use of *solely* before the subject of each object property. In the table, *I* is used to represent the object property. In the actual questions, *is_nearest_neighbour_of* was used.

Table 17 – Questions employing inverse functional object properties

	axioms (I = *is_nearest_neighbour_of)*	**putative conclusion**	**validity**	**relational complexity**
variant 1				
7	r I s; t I s; v I r; w I t	v SameAs w	valid	3,2,3
8	r I s; t I v; r DifferentFrom t; s I w; x I z; v SameAs x	w DifferentFrom z	valid	4,2,4
variant 2				
7	solely r I s; solely t I s; solely v I r; solely w I t	v SameAs w	valid	3,2,3
8	solely r I s; solely t I v; r DifferentFrom t; solely s I w;	w DifferentFrom z	valid	4,2,4

[13] A caveat applies here. In these comparisons, for both studies we are using six questions out of a total of eight in the section. For study 2 the lack of randomization in the question order could introduce a bias. However, this is unlikely to be appreciable. The same caveat applies to other comparisons in Section 7.

	solely x I z; v SameAs x			

Table 18 shows the accuracy and response time data for these questions. For question 7, the accuracy for variant 2 was greater, but not significantly so ($p = 0.651$, Fisher's Exact Test, two-way). For question 8, the percentage of correct responses was the same in both variants.

For question 7, the response time was greater for variant 2 than for variant 1. For question 8, the response time was less for variant 2. However, in neither case was the difference significant (question 7: $t(26.423) = 1.2631$, $p = 0.218$; question 8: $t(19.906) = 0.57643$, $p = 0.571$).

Table 18 – Questions employing inverse functional object properties - accuracy and response time data

	variant 1 – without solely; N = 15		**variant 2** – with solely; N = 15	
	% corr	mean time (SD) - secs	% corr	mean time (SD) - secs
question 7	73%	38 (18)	87%	48 (23)
question 8	73%	105 (92)	73%	90 (43)

7.2 Complement, intersection and union

Study 1 noted the difficulty experienced with complemented intersection. Subsection 5.3.1 suggested that this might be the result of a failure to create all three of the mental models required. This failure could, in part, be caused by the ambiguity surrounding the word *and*, as discussed in subsection 3.2. It was hypothesized that the use of the keyword *intersection* would reduce this problem:

> H3.3 The use of the keyword *intersection* in place of *and* will improve participant performance for complemented intersection.

A number of the questions in study 2 made use of *and not* to create exceptions. It was hypothesized that the use of the keyword *except* would improve reasoning:

> H3.4 The use of *except* in place of *and not* will improve participant performance.

Table 19 shows the questions for variant 1, which were used to investigate these two hypotheses. H3.3 was investigated by comparing question 1 in variant 1 with a comparable question in study 1, specifically question 2 in Table 4. For consistency, *or* was replaced with *union*, and the effect of this change was investigated by comparing question 2 in variant 1 with question 4 in Table 4. In study 1, both these questions were based on an ontology pattern. For study 3 the essential features of the questions have been extracted out to make them standalone.

H3.4 was investigated by comparing questions 3 to 8 with comparable questions from study 2, specifically, questions 1, 2, 3, 4, 9 and 10 in Table 7. These questions were chosen to represent all three levels of complexity. In Table 7 there are six questions of medium complexity; questions 3 and 4 in Table 7 were chosen to generate questions for this study because these questions made the greatest use of *and not*. As in the previous section, the names of the entities were changed.

Table 19 Boolean concept constructor questions – variant 1

	V / NV	axioms	putative conclusion
Derived from study 1			
1	NV	UNIVERSE DisjointUnionOf W, X, Y, Z; a Type UNIVERSE; a Type not (W intersection Y);	a Type (X union Z)
2	V	UNIVERSE DisjointUnionOf W, X, Y, Z; a Type UNIVERSE; a Type not (W union Y);	a Type (X union Z)
Derived from study 2			
3	V	W EquivalentTo ((UNIVERSE except X) except Y); UNIVERSE DisjointUnionOf X, Y, Z	W EquivalentTo Z
4	V	W EquivalentTo (UNIVERSE except (X union Y)); UNIVERSE DisjointUnionOf X, Y, Z	W EquivalentTo Z
5	V	W EquivalentTo (UNIVERSE except (X except X_1)); UNIVERSE DisjointUnionOf X, Y; X DisjointUnionOf X_1, X_2	W EquivalentTo (Y union X_1)
6	NV	As for question 5	W EquivalentTo Y
7	V	W EquivalentTo (UNIVERSE except (X except (X_1 except X_1_A))); UNIVERSE DisjointUnionOf X, Y; X DisjointUnionOf X_1, X_2; X_1 DisjointUnionOf X_1_A, X_1_B	W EquivalentTo (Y union X_1_B)
8	NV	As for question 7	W EquivalentTo X_1_B

MOS uses infix notation for intersection and union. However, earlier DL syntaxes used prefix notation. Variant 2 uses prefix notation to enable a comparison between infix and prefix. This is discussed in Warren et al. (2017), where it is shown that there was no significant difference between the two approaches. A caveat needs to be stated here. The participants were almost all experienced in various aspects of computer science, as is demonstrated by the data in subsection 7.5. It may be that the infix notation, familiar from everyday language, would be of value to domain experts with limited experience of computer science.

Table 20 shows the accuracy and response time data for questions 1 and 2, and Table 21 shows the data for questions 3 to 8. For comparison, the relevant data from the previous two studies are also shown.

Table 20 Boolean concept constructor questions – study 3: questions 1 and 2

study 1 from Table 4			study 3				
				variant 1		variant 2	
	% age correct N = 12	mean resp. time (s.d) - secs N = 12		% age correct N = 15	mean resp. time (s.d) - secs N = 15	% age correct N = 15	mean resp. time (s.d) - secs N = 15
2	25%	75 (48)	1	80%	53 (41)	67%	47 (16)
4	92%	44 (19)	2	87%	42 (28)	100%	39 (25)

Table 21 Boolean concept constructor questions – study 3: questions 3 to 8

study 2 from Table 8			study 3				
				variant 1		variant 2	
	% age correct N = 28	mean resp. time (s.d) – secs N = 24		% age correct N = 15	mean resp. time (s.d) – secs N = 15	% age correct N = 15	mean resp. time (s.d) – secs N = 15
1	82%	39 (26)	3	100%	39 (25)	100%	47 (30)
2	86%	43 (29)	4	100%	35 (26)	93%	46 (35)
3	61%	96 (56)	5	53%	61 (37)	53%	65 (38)
4	64%	105 (78)	6	100%	44 (25)	73%	82 (57)
9	54%	90 (48)	7	60%	97 (75)	40%	156 (126)
10	68%	94 (47)	8	80%	88 (60)	60%	93 (51)
Mean	69%	78 (56)	Mean	82%	61 (50)	70%	82 (74)

7.2.1 Replacing *and* with *intersection* – hypothesis H3.3

To investigate the effect of replacing *and* with *intersection*, a comparison was made between question 2 of study 1 and question 1 of study 3, variant 1. The latter was answered significantly more accurately than the former ($p = 0.007$, Fisher Exact Test, two-sided). However, there was no significant difference in response time ($t(23.988) = 1.6031$, $p = 0.122$).

A comparison of question 4 of study 1 with question 2 of study 3 variant 1 revealed that the replacement of *or* with *union* made no significant difference to accuracy ($p = 1$, Fisher Exact Test, two-sided). This is to be expected given the high proportion of correct responses for this question in study 2. Nor was there any significant difference in response time ($t(24.433) = 0.5372$, $p = 0.596$).

7.2.2 Replacing *and not* with *except* – hypothesis H3.4

To investigate the effect of replacing *and not* with *except* a comparison needs to be made between questions 3 to 8 of variant 1 in study 3 and the corresponding questions in study 2. Inspection of Table 21 indicates that, taking the aggregate for these questions, variant 1 of study 3 was answered more accurately than the study 2 questions. This difference was significant ($p = 0.026$, Fisher Exact Test, two-sided). When the aggregated response times are compared, the study 3 variant 1 questions can be seen to be answered significantly faster ($t(201.95) = 2.4906$, $p = 0.014$).

7.3 Complement and restrictions

Subsection 6.4 described the difficulties which study 2 participants experienced with the universal and existential restrictions, and related these difficulties in part to a failure to form both the required mental models. Alternative keywords to *only* and *some* might draw attention to the less dominant mental model and improve participant performance. In particular, it was hypothesized that:

H3.5 The use of *noneOrOnly* in place of *only* and *including* in place of *some* will lead to improved participant performance.

In addition, in study 2, some questions included the construct *not <object property> some*, e.g. two questions in Table 9 included the axiom *X SubClassOf not (has_child some MALE)*. However, *not ... some* is not a natural English construct; a more natural construct is *not ... any*. The use of the latter construct might aid comprehension and reasoning. This leads to the hypothesis:

H3.6 The use of *any* to indicate the existential restriction, when the corresponding object property is preceded by a complement, will improve performance.

To test these hypotheses, the questions from Table 9 involving complement and restrictions were modified as in Table 22. Variant 1 was constructed by substituting *noneOrOnly* for *only* and *including* for *some*. Variant 2 was constructed similarly, except that *not ... any* was substituted for *not ... some*; note that this only affects questions 3 and 7. Table 23 shows the accuracy and mean response times for the questions, and for the comparable questions in study 2.

Table 22 Complement and restriction questions

	first axiom	**second axiom**	**validity**
1	`A SubClassOf has_child including (not FEMALE)`	B SubClassOf has_child noneOrOnly FEMALE	valid
2	A SubClassOf has_child noneOrOnly (not FEMALE)		not valid
3	**variant 1**		not valid
	A SubClassOf not (has_child including FEMALE)		
	variant 2		
	A SubClassOf not (has_child any FEMALE)		
4	`A SubClassOf not (has_child noneOrOnly FEMALE)`		valid
5	`A SubClassOf has_child including (not FEMALE)`	B SubClassOf has_child including FEMALE	not valid
6	A SubClassOf has_child noneOrOnly (not FEMALE)		valid
7	**variant 1**		valid
	A SubClassOf not (has_child including FEMALE)		
	variant 2		
	A SubClassOf not (has_child any FEMALE)		
8	`A SubClassOf not (has_child noneOrOnly FEMALE)`		not valid

N.B. the putative conclusion in each case was A DisjointWith B.

Table 23 Complement and restriction questions – accuracy and response times

	study 2 from Table 10		**study 3 both variants**		**study 3 variant 1**		**study 3 variant 2**	
	only, some		*noneOrOnly, including*					
					not ... including		*not ... any*	
	%age correct N = 28	mean resp. time (s.d.) - secs N = 24	%age correct N = 30	mean resp. time (s.d.) - secs N = 30	%age correct N = 15	mean resp. time (s.d.) - secs N = 15	%age correct N = 15	mean resp. time (s.d.) - secs N = 15
1	61%	52 (39)	80%	42 (33)				
2	50%	33 (18)	73%	29 (20)				
3	68%	45 (22)	70%	69 (127)	67%	55 (49)	73%	84 (175)
4	75%	43 (25)	90%	41 (32)				
5	64%	41 (30)	70%	30 (21)				
6	50%	44 (40)	70%	33 (33)				
7	79%	43 (37)	80%	29 (16)	73%	26 (14)	87%	33 (18)
8	68%	60 (37)	67%	38 (24)				
Exc Q3 and Q7	61%	45 (33)	75%	35 (28)				
Q3 and Q7	73%	44 (30)	75%	49 (92)	70%	40 (38)	80%	58 (125)

7.3.1 *noneOrOnly* and *including* – hypothesis H3.5

To avoid the confounding effect of the use of *not ... any* for questions 3 and 7 in variant 2, a comparison of studies 2 and 3 used the remaining six questions. For these six questions, using data from variant 1 and variant 2, study 3 questions were answered significantly more accurately than study 2 questions ($p = 0.008$, Fisher's Exact Test, two-sided) and also significantly more quickly ($t(284.57) = 2.7897$, $p = 0.006$).

7.3.2 *not … any* – hypothesis H3.6

It is not possible to compare the effect of using *not ... some* in study 2 with *not ... any* in study 3, variant 2, because of the confounding effect of the use of *noneOrOnly* or *including* in the second axiom. Therefore, the comparison made here is between *not ... including* in study 3 variant 1 and *not ... any* in variant 2. Table 23 shows that, for questions 3 and 7 the accuracy was greater for variant 2, which used *not ... any*. However, the difference was not significant ($p = 0.552$, Fisher's Exact Test, two-sided). Table 23 shows that, for these two questions, the response time was greater for variant 2. However, again this difference was not significant ($t(57.832) = 0.79496$, $p = 0.430$).

7.4 Nested restrictions

The questions in study 2 with nested restrictions offer another opportunity to investigate hypotheses H3.5 and H3.6. Table 24 shows the analogous questions used in variant 1 of study 3, where *only* has been replaced with *noneOrOnly* and *some* with *including*. Comparison of variant 1 with study 2 enables the further investigation of H3.5. For variant 2, *only* and *some* have also been replaced with *noneOrOnly* and *including*; except that for questions 5 to 8, in the final axiom *not ... any* is used in place of *not ... including*. This enables a further investigation of hypothesis H3.6.

Table 24 Nested restriction questions – study 3; variant 1

	first axiom(s)	final axiom	valid
1	A SubClassOf (has_child including (has_child including MALE))	x has_child y; y Type has_child including (not MALE)	not valid
2	A SubClassOf has_child including B; B EquivalentTo has_child noneOrOnly MALE		not valid
3	A SubClassOf (has_child noneOrOnly (has_child including MALE))		not valid
4	A SubClassOf has_child noneOrOnly B; B EquivalentTo has_child noneOrOnly MALE		valid
5	A SubClassOf has_child including B; B EquivalentTo has_child including MALE	x has_child y; y Type (not (has_child including MALE))	not valid
6	A SubClassOf (has_child including (noneOrOnly MALE))		not valid
7	A SubClassOf has_child noneOrOnly B; B EquivalentTo has_child including MALE		valid
8	A SubClassOf (has_child noneOrOnly (has_child noneOrOnly MALE))		not valid

N.B. the putative conclusion in each case was x Type (not A).

The final hypothesis is not directly concerned with the main theme of this study, but nevertheless is concerned with how MOS can best be written to aid comprehension. Inspection of the response time data for this section in study 2 suggested that participants were taking longer overall to answer those questions which employed a named class to link the first two axioms, compared with those questions where the first axiom included two restrictions. However, the fact that the two formats were used for different questions made a rigorous comparison impossible. As can be seen from Table 24, in study 3 variant 1 used the same format as study 2; enabling a proper investigation of hypotheses H3.5. Questions 5 to 8 of variant 2 also used the same format as study 2, and hence of variant 1, enabling a proper investigation of hypothesis H3.6. However, questions 1 to 4 of variant 2 used the alternative format to study 2 and variant 1. For example, question 1 in variant 1 (Table 24) used an anonymous class, whereas question 1 of variant 2 (not shown) used a named class. Conversely, question 2 of variant 1 used a named class; whereas question 2 of variant 2 used an anonymous class. This enabled a comparison between the two formats and an investigation of the hypothesis:

> H3.7 There will be a difference in reasoning performance between the equivalent use of a named class and an anonymous class.

Table 25 shows the accuracy and mean response times for both variants of study 3, and for comparison repeats the data from study 2 shown in Table 12. Note that the aggregate data for variant 2 is not shown. The use of *not … including* in questions 1 to 4 and *not … any* in questions 5 to 8 has the consequence that the aggregate data for variant 2 are not useful for comparison.

Table 25 Nested restriction questions – accuracy and response time

	study 2 from Table 12		study 3 var 1		study 3 var 2	
	only, some		*noneOrOnly, including*			
	% age correct N = 28	mean resp. time (s.d.) – secs N = 24	% age correct N = 15	mean resp. time (s.d.) – secs N = 15	% age correct N = 15	mean resp. time (s.d.) – secs N = 15
1	71%	69 (45)	80%	47 (30)	60%	73 (59)
2	57%	79 (53)	40%	65 (20)	67%	68 (41)
3	71%	63 (43)	60%	54 (31)	53%	79 (46)
4	57%	63 (39)	53%	100(87)	47%	66 (49)
					not ... any	
5	54%	88 (62)	40%	64 (30)	67%	74 (67)
6	64%	73 (45)	73%	85 (82)	73%	99 (90)
7	71%	80 (36)	53%	64 (39)	47%	97(102)
8	50%	55 (30)	60%	83 (35)	53%	63 (37)
Mean for Q5 to 8	60%	74 (46)	57%	74 (51)	60%	83 (77)
Mean for all questions	62%	71 (45)	58%	70 (51)		

7.4.1 *noneOrOnly* and *including* – hypothesis H3.5

To investigate the effect of replacing *some* and *only* with *including* and *noneOrOnly* requires a comparison of study 3 variant 1 with the analogous section in study 2. Inspection of Table 25 shows that the overall accuracy for the former was actually less than for the latter, although the difference was not significant ($p = 0.420$, Fisher's Exact Test, two-sided). There was also no significant difference in response times ($t(288.79) = 0.40724$, $p = 0.684$). This contrasts with the results of subsection 7.3.1 where, for those questions, there was a significant effect on accuracy and response time.

7.4.2 *not … any* – hypothesis H3.6

To investigate the effect of replacing *not … including* with *not … any* requires a comparison of questions 5 to 8 in the two variants of study 3. Inspection of Table 25 shows that, considering questions 5 to 8, the overall accuracy was slightly greater for variant 2 than variant 1, although the difference was not significant ($p = 0.853$, Fisher's Exact Test, two-sided). There was also no significant difference in response times ($t(106.86) = 0.19408$, $p = 0.847$). This result is consistent with that of subsection 7.3.2.

7.4.3 named versus anonymous classes – hypothesis H3.7

Table 26 shows the data from questions 1 to 4, aggregated over the questions using a named class and over the questions where the named class was replaced with an anonymous class. Note that each participant answered two of the four questions with a named class and the other two with an anonymous class.

Table 26 Named and anonymous classes – percentage correct and response times

named class; N = 15		**anonymous class**; N = 15	
% corr	mean time (s.d.) - secs	% corr	mean time (s.d.) - secs
52%	79 (58)	63%	59 (39)

Inspection of Table 26 shows that the questions using an anonymous class were answered more accurately. However, the difference was not significant ($p = 0.268$, Fisher Exact Test, two-sided). The questions using an anonymous class were answered more quickly than the questions using a named class, and in this case the difference was significant ($t(117.99) = 2.5387$, $p = 0.012$).

It is only possible to speculate on what causes this significant difference in response times. Possibly, where two statements are used, with a named class in common, two mental models are formed, which then have to be merged. Whereas, where one statement is used, with the named class being replaced by an anonymous class, then one mental model is created as the statement is being parsed. Another possibility is that storing and retrieving the name, and information associated with the name, takes more cognitive effort and hence time. It may be that the use of a meaningless name increases this effect, and that a meaningful name (e.g. *parentOfSomeMale*) would reduce the effect.

7.5 Effect of participants' background knowledge

As in the previous studies, participants were asked to rate their knowledge of formal logic and of OWL or another DL formalism. For this study there is the additional complication that difference between the two variants may act as a confounding factor. Hence, each variant has been treated separately. Table 27 shows, for each variant, how the respondents divided between the various categories of knowledge of logic, with the mean accuracy and time in each category. Table 28 provides similar information for knowledge of OWL or another DL formalism. Inspection of both tables suggest, *prima facie*, that knowledge of logic and knowledge of DL both correlate with increased accuracy and reduced time. However, there are only three significant results. In variant 1, knowledge of formal logic correlates significantly with speed of answering the questions. In variant 2, knowledge of logic and knowledge of DL formalism both correlate significantly with accuracy.

Table 27 Participant performance by knowledge of formal logic

	Variant 1				**Variant 2**		
	N	%age correct	mean time (s.d.) - secs		N	%age correct	mean time (s.d.) - secs
No knowledge	1	69%	2722 (NA)		0	NA	NA (NA)
A little knowledge	3	60%	2484 (859)		4	59%	2463 (526)
some knowledge	9	75%	1911 (374)		5	74%	2571 (697)
expert knowledge	2	73%	1524 (26)		6	79%	2043 (525)
Spearman's rank correlation (one-sided)		$\rho = 0.362$ $p = 0.092$	$\rho = -0.661$ $p = 0.004$			$\rho = 0.585$ $p = 0.011$	$\rho = -0.277$ $p = 0.159$

Table 28 Participant performance by knowledge of OWL or another DL formalism

	Variant 1				Variant 2		
	N	%age correct	mean time (s.d.) - secs		N	%age correct	mean time (s.d.) - secs
No knowledge	3	82%	2077 (607)		1	66%	3073 (NA)
A little knowledge	6	66%	2116 (703)		8	68%	2391 (592)
some knowledge	5	69%	1997 (444)		3	70%	2339 (539)
expert knowledge	1	81%	1506 (NA)		3	89%	1913 (660)
Spearman's rank correlation (one-sided)		$\rho = -0.187$ $p = 0.748$[14]	$\rho = -0.223$ $p = 0.212$			$\rho = 0.483$ $p = 0.034$	$\rho = -0.350$ $p = 0.101$

An additional question is whether knowledge of formal logic or DL has a different effect on the different sections, which might require different kinds of expertise. To investigate this in the case of accuracy, a logistic analysis of deviance was undertaken. Four factors were used: knowledge of logic, knowledge of DL, section, and variant. The last factor was included because, as can be seen from Tables 27 and 28, the distribution of expertise was different for the two variants. Effects due to differences in expertise need to be separated from effects due to difference between the variants. The analysis included the effect of pairwise interactions. A difficulty with such an analysis is that, because the data is unbalanced, the results of the analysis of deviance will depend upon the order of specification of the factors. The analysis was executed in the order: knowledge of logic, knowledge of DL, section and variant; and also with knowledge of logic and knowledge of DL interchanged. In both cases, the two kinds of expertise and the section were significant ($p < 0.001$ for all three factors). However, neither the variant nor any of the pairwise interactions was significant ($p > 0.1$ in all cases).

A parallel analysis of variance applied to the log(time) data revealed a more complex situation. When the factor knowledge of logic preceded knowledge of DL, then the time was significantly dependent on knowledge of logic ($p = 0.031$), section ($p = 3.25 \times 10^{-8}$), variant ($p = 0.015$), but not knowledge of DL ($p = 0.665$). There were two interaction effects: between knowledge of logic and variant ($p = 0.036$) and knowledge of DL and variant ($p = 0.013$). When the order of knowledge of logic and DL was interchanged, neither of these two factors were significant. The only significant factors were: section ($p = 3.25 \times 10^{-8}$), variant ($p = 0.015$), and the interaction between knowledge of logic and variant ($p = 0.000717$). In neither case was there any significant interaction between prior knowledge and section.

To summarise, it appears that prior knowledge is beneficial, and that such knowledge does not differentially affect the different sections.

7.6 Conclusions from study 3

Table 29 summarises the findings from study 3. The study demonstrated that the use of a different, or in one case an additional, keyword could significantly affect performance. The use of *solely* to indicate the direction of uniqueness with functional properties significantly reduced response time for the valid questions, although it had no significant effect for inverse functional properties. In one question section, described in section 7.3, the replacement of *some* and *only* with *including* and *noneOrOnly* significantly increased accuracy and significantly reduced response time. However, in the section with nested restrictions,

[14] Note that the correlation coefficient is, counter-intuitively, negative. However, as with other correlations between knowledge and accuracy, the p-value has been calculated on the prior assumption of a positive correlation.

described in section 7.4, these changes made no significant difference, neither to accuracy nor response time. In study 2 the nested restrictions questions were answered significantly more slowly than the other section involving restrictions. It may be that these questions are too difficult for most participants to answer correctly, whatever keywords are used.

Table 29 Summary of study 3 findings

	accuracy	**response time**
H3.1 Use of *solely* in functional object property.	No significant effect.	Significantly reduced response time for the valid questions.
H3.2 Use of *solely* in inverse functional object property.	No significant effect.	No significant effect.
H3.3 Use of *intersection* in place of *and* with complemented intersection.	Significant increase in accuracy.	No significant effect.
H3.4 Use of *except* in place of *and not*.	Significantly increased accuracy.	Significantly reduced response time.
H3.5 Use of *noneOrOnly* in place of *only* and *including* in place of *some*.	Significant increase in accuracy for questions without nested restrictions, but no significant effect for questions with nested restrictions.	Significant reduction in response time for questions without nested restrictions, but no significant effect for questions with nested restrictions.
H3.6 Use of *not … any* in place of *not … some.*	No significant effect.	No significant effect.
H3.7 Use of a named class versus an anonymous class.	No significant effect.	Reasoning with anonymous class significantly faster than with named class.

8 Implications and future work

This section considers some of the implications of the work and proposes some future research directions. Subsection 8.1 makes some proposals for improving the comprehension of DLs. Subsection 8.2 discusses the role of theory. Subsection 8.3 looks at the role of natural language and also the relationship between FOL and DLs. Subsection 8.4 discusses some future research directions and subsection 8.5 makes some final comments.

8.1 Proposals

These studies lead to proposals in three areas: training; support systems; and the syntax of MOS.

8.1.1 Training

There are four areas in particular which require emphasis in training. Firstly, it needs to be made clear that property characteristics are not necessarily inherited by subproperties. The most effective way to do this may be to provide examples where inheritance fails. For

example, *siblingOf* is symmetric, but its two subproperties *brotherOf* and *sisterOf* are not; *descendantOf* is transitive, but *sonOf* and *daughterOf* are not. For more advanced students, it may be desirable to explain which characteristics are inherited, and why. The difference between inherited and non-inherited characteristics is discussed in Warren (2017; Chapter 2)[15].

Secondly, the difference between functional and inverse functional properties needs to be stressed, so that students are clear that in the former case it is the object which is unique, and in the latter case it is the subject. For functional properties, the analogy with elementary mathematics may be helpful. *f(x)* always evaluates to one number, although *f(x) = k* may have several solutions. Inverse functional properties can then be considered as the reverse of functional.

Thirdly, attention needs to be given to De Morgan's laws. The rule is that, when the complement moves inside or outside the brackets, then the operator switches between intersection and union. This can be illustrated with the use of Venn diagrams.

Fourthly, related duality laws for restrictions need to be introduced. The rule is analogous to that for De Morgan's laws. When the complement moves inside or outside the restriction, then the restriction switches between existential and universal. A diagrammatic approach can be used to illustrate this. Figure 6 illustrates that *some* and *only not* are complementary, and hence that of *not ... some* and *only not* are equivalent. Alternatively, the rule can be illustrated by analogy with the following, or similar, sentences:

(1) John has some non-male children.
(2) John does not have some male children.
(3) John has only non-male children.
(4) John does not have only male children.

At first glance one might think that (1) is equivalent to (2). They both use *some* and the negation has simply been moved to a different place in the sentence. Similarly, one might think that (3) is equivalent to (4). More careful thought reveals that (1) is equivalent to (4) and (2) is equivalent to (3), where the latter equivalence takes into account the possibility that John has no children at all.

[15] In brief, the distinction is between a characteristic which implies the validity of a property relationship and a characteristic which implies the falsehood of a property relationship. For example, compare a symmetric property, *S*, and an asymmetric property, *A*. For the former, we have:

$$a\ S\ b \Rightarrow b\ S\ a$$

For the latter we have:

$$a\ A\ b \Rightarrow \neg\ b\ A\ a$$

The symmetric characteristic is not inherited by a subproperty of *S*, e.g. *P*, since given *a P b* we cannot demonstrate that *b P a* is true. All we can say is:

$$a\ P\ b \Rightarrow a\ S\ b \Rightarrow b\ S\ a$$

However, the asymmetric characteristic is inherited by a subproperty of *A*, e.g. *Q*, since if both *a Q b* and *b Q a*, then we have *a A b* and *b A a*, which contradicts the fact that *A* is asymmetric.

	non-male child(ren)	no non-male children
male child(ren)		
no male children		

Figure 6. Illustrating the complementary nature of *some male children* and *only non-male children*; the former is shown with horizontal stripes and the latter with vertical. Hence, *not some male children* is equivalent to *only non-male children*.

More generally, attention should be given to tracking the alternative mental models, e.g. associated with restrictions. Teaching students initially to externalize these, e.g. diagrammatically, would reinforce an awareness of the various possibilities.

8.1.2 Support systems

In some situations it might be useful to automatically generate alternative representations of the same semantics, so that the ontologist can see both versions and use whichever is the more helpful. This might be particularly useful where one of the duality laws for restrictions can be invoked to give an alternative representation. The manipulation of statements with Boolean constructors, e.g. by using De Morgan's laws, could also be helpful in generating alternative, more comprehensible, representations.

Visualizations of DL statements could be useful. Stapleton et al. (2013; 2014) illustrate how to represent common DL axioms using concept diagrams, so as to support reasoning about those statements. However, rather than being applied to a whole ontology, visualization could also be applied to a subset of the ontology axioms, e.g. the justification for an entailment. This could be done at the request of the ontologist, when difficulties in comprehension are being experienced. Attention should also be given to the choice of names; explaining how particular choices can support human reasoning and others lead to unwanted implicatures, e.g. of property characteristics.

8.1.3 Syntax

Replacing *and* and *or* by *intersection* and *union* would be helpful, as would replacing *and ... not* by *except*. Using the additional keyword *solely* to indicate the direction of uniqueness in a functional object property would also assist ontologists. Something similar might be useful with inverse functional properties, although as discussed later, more work is required to determine a suitable keyword and the optimum position. Replacing *some* and *only* by *including* and *noneOrOnly* might also help, although there may be better alternative keywords. Finally, whilst there may be sound design reasons for using named classes rather than anonymous classes in some situations, this may hinder comprehension.

8.2 The role of theory

The research set out to investigate whether theory, from psychology and the philosophy of language, could help to explain the difficulties experienced in reasoning about DLs. Theory has served two purposes. Firstly, it has been used to understand these difficulties, and hence to help develop strategies for mitigating them. This aspect has been present in the research relating to Boolean concept constructors and restrictions. Secondly, psychological theory has

been used to provide a measure of complexity, and this has supported research into functional object properties.

The discussion of theories of reasoning in section 3 began with the rule-based approach. The reason for this was partly historical; the rule-based theory came first. Moreover, it is necessary to understand this theory to properly appreciate the model-based theory which came later and arose in opposition to it. Whilst the model-based theory has been used to offer a possible explanation for some of our experimental results, this is not directly the case for the rule-based theory. No recourse has been made to the kind of logic rules investigated by Rips (1983). However, rule-based reasoning is an example of syntactic reasoning, and there were cases where participants appeared to be reasoning syntactically. Syntactic reasoning offers a shortcut over building models, e.g. when syntax identifies obviously complementary classes. When that shortcut is available, human reasoners appear often to take it. Sometimes the shortcut is deceptive and leads to error, e.g. when failing to distinguish between *some not* and *not … some*. The choice between the two modes of reasoning may sometimes be a matter of personal preference. Ford (1995) has observed that some people exhibit a preference either for syntactic or spatial reasoning, and the latter could be interpreted as a kind of model-building.

Both theoretical approaches provide measures of complexity, e.g. related to the number and difficulty of the reasoning steps and the number and complexity of the mental models. However, in the work with functional object properties, RC theory offered a natural measure of complexity which was used to make a controlled comparison between the difficulty of reasoning with functional and transitive object properties.

Turning to the philosophy of language, the concept of the implicature helped understand how the ideas associated with the natural language keywords *only* and *some* can lead human reasoners astray when building mental models. The difficulties associated with the use of natural language are discussed in detail in the next subsection.

8.3 The role of natural language

The previous subsection noted the problems associated with the use of *only* and *some*. Another problem arose with the use of *and*, specifically when used together with complement. Study 1 showed that complemented intersection was a particular problem in MOS. Study 3 then showed that accuracy could be significantly improved by replacing the ambiguous keyword *and* with the keyword *intersection*.

Section 3.2 discussed in detail the ambiguities inherent in the use of *and* and *or*, which in everyday language are frequently resolved by context. It was shown that, starting from the use of *and* to represent the conjunction of propositions, *and* could in one context be interpreted as meaning union, and in another context as meaning intersection. This raises a particular problem with DLs. Baader and Nutt (2003) describe DLs as "a family of knowledge representation formalisms … equipped with a formal, logic-based semantics". DLs are logic-based; their theoretical development is related to theories of logic. However, from the standpoint of a domain expert using a DL, e.g. OWL, there is a fundamental difference between DLs and logic. Logic, e.g. FOL, is concerned with propositions; DLs are concerned with classes. In everyday English, whilst *and* is unambiguous when used with propositions, conveying the sense of conjunction, when used with classes it can mean either intersection or union, and indeed the latter may be the more common meaning.

Recent work (Alharbi et al., 2017) demonstrated that novices achieved as much accuracy with the German DL syntax as with MOS. In the reported study, out of a total of 18 tasks, there was no significant difference in accuracy for 16 of them. For the two where there was a significant difference, the more accurate results were achieved with German DL. It needs to be remembered that MOS is not natural English. It is a formal language with keywords, the meaning of which need to be learned, just as do the symbols of logic.

The purpose of this discussion is not to reject the use of natural language keywords altogether. Indeed, section 7 reported favourably on the use of *except*. The purpose is rather to draw attention to the care which needs to be taken when using natural language in computer interaction; and the consideration which needs to be given to the meaning which words convey in everyday use.

8.4 Future research

The work described here leaves some research questions unresolved, and these are examined in subsection 8.4.1. Subsection 8.4.2 then discusses some alternative methodologies which would complement our approach. Subsection 8.4.3 discusses how alternative theories of reasoning, developed more recently, could be relevant to how people interact with information. Finally, subsection 8.4.4 expands the research horizon by discussing alternative application areas.

8.4.1 Unresolved questions

The results of study 3 raise some immediate areas for investigation. One open questions is whether alternative keywords to *including* and *noneOrOnly* would improve performance with the existential and universal restrictions. For example, Krötzsch et al. (2012), discussing the universal restriction, equate the expression $\forall parentOf.Female$ to the English phrases "no children other than female ones" and "no children that are not female". This suggests *noneOtherThan* or *noneNot* as alternatives to *only*. Discussing the existential restriction, they equate $\exists parentOf.Female$ to "individuals that are parents of at least one female individual". This suggests *atLeastOne* as an alternative to *some*.

Improving performance with inverse functional properties is another area for research. In study 3, the use of *solely* immediately before the subject was not successful. It may be that it is not sufficiently clear that *solely* refers to the subject rather than the object. An alternative keyword, e.g. *alone*, immediately after the subject and before the object property might better convey the subject's uniqueness.

The use of object subproperties was examined in study 1 but not further investigated. The keyword *SubPropertyOf* may be misleading because the prefix *sub* conveys the implicature of inheritance, by analogy with object-oriented programming and everyday English usage. If *S* is a subproperty of *P*, then this means that *a S b* implies *a P b*. An alternative way of specifying this might be to use a keyword which avoids the implicature of inheritance, e.g. *S implies P*. Looked at this way, a subproperty hierarchy is no more than a chain of inferences. Indeed, in everyday English, whilst we talk about subclasses and subsets, we rarely talk about subproperties; e.g. 'being a sister is a subproperty of being a sibling' sounds much less natural than 'being a sister implies being a sibling'. To summarise: in DL, the use of the

word ‘subproperty’ conveys the implicature of an analogy between subproperty and subclass which does not exist.

Inevitably, the studies could not examine all the potentially interesting DL constructs. One such is the use of object properties defined to be inverse. This was relatively often used according to the study of Power and Third (2010) and the survey conducted by the current authors (Warren, et al., 2014b).

The work by Alharbi et al. (2017) challenges the conventional view that non-logicians are better served by English-based languages rather than the notation of logic. Certainly, the use of a logical notation removes the problems of implicatures associated with keywords, and leaves only those problems which are inherent in DLs. Alharbi et al. (2017) themselves comment on the need for more work to compare German DL with MOS. It would also be useful to compare German DL with proposed improvements to MOS. It might be that judicious choice of keywords could improve on German DL.

Recent work by Sarker et al. (2017) has suggested that the use of rules, specifically from the Semantic Web Rule Language (SWRL)[16] can lead to significant improvements in modelling time and accuracy compared with an equivalent OWL representation. There are difficulties, however. Not all SWRL rules can be represented in DL; and, whilst SWRL itself is decidable, the combination of SWRL rules with decidable subsets of OWL can lead to undecidability. Nevertheless, Sarker et al. (2017) suggest “the development of a full-blown rule syntax” for OWL. This goes further than the work reported here, in offering the possibility not just of changes to syntax, but of a different modelling paradigm, although completely compatible with OWL.

8.4.2 Alternative methodologies

The three studies were concerned with comprehension of, and reasoning with, DL constructs identified as being commonly used in ontologies. Given that one of the motivations of the work was the problems experienced by ontologists in understanding the reasoning chain from justification to entailment, an alternative approach would be to identify the deduction rules commonly used in such reasoning chains. Nguyen et al. (2012) did this by computing entailment-justification pairs. Another strategy would be to ask ontologists to provide entailment-justification pairs which they had found difficult.

Whilst the emphasis in this work has been on comprehension and reasoning, modelling may present different problems. The syntax changes proposed in Section 7 may have an effect on modelling performance, and this also needs investigation through controlled experiments.

Controlled experiments offer great insights and are relatively efficient in the use of the experimenter’s time, at the expense of a certain unnaturalness. A useful complement would come through greater interaction with users, in interviews and focus groups, and by observation of them in their work.

8.4.3 Alternative theories of reasoning

[16] https://www.w3.org/Submission/SWRL/

The theories of reasoning used here were chosen in part because of their fundamental position in reasoning research. They were the first to offer insight into reasoning and are still of interest in the psychological research community. However, two further developments are worth commenting on, since they offer possible additional relevant insights: probabilistic models and models of human non-monotonic reasoning. These models of reasoning may be particularly relevant to the presentation of information from the Web, where people need to cope with incomplete, incorrect and sometimes inconsistent information, and think in terms of probabilities.

For an early discussion of probabilistic models, see Oaksford and Chater (2001), who apply the probabilistic approach to conditional inference, Wason's selection task and syllogistic reasoning. The use of probabilistic reasoning makes it theoretically possible to unify deductive with inductive reasoning. Deductive reasoning can be seen as reasoning about statements with an associated probability equal to one. Inductive reasoning can be seen as reasoning about statements with associated probability potentially less than one.

Non-monotonic reasoning has been observed in humans for some time. Byrne (1989) gives an example based on the following three statements:

1. If she has an essay to write then she will study late in the library.
2. If the library stays open then she will study late in the library.
3. She has an essay to write.

Statements (1) and (3) should lead to the conclusion that "she will study late in the library". However, the presence of statement (2) resulted in 38% of study participants making this conclusion and 62% concluding that "she may or may not study late in the library". Ragni et al. (2016) consider non-monotonic logics and conclude that some "seem to be adequate to describe human commonsense reasoning". Ideas from non-monotonic logics may be relevant to researching how humans reason when a statement (e.g. 2 above) tends to suppress a conclusion from other statements (e.g. 1 and 3 above).

8.4.4 Expanding the research horizon

The work reported here has been concerned with applying theories from cognitive psychology and philosophy of language to one particular area, that of DLs. However, insights from these disciplines may be applicable to other areas of knowledge representation and also knowledge querying.

One such area is that of database interaction. Database usability has been a subject of research since the early days of databases, and some of the discussion echoes some of our discussion about DLs. As an early example, Thomas and Gould (1975) looked at a query system using a formal language. One difficulty they identified was with universal quantification. Conversely, they found little difficulty with conjunction and disjunction. However, in an acknowledgement that disjunction and conjunction might prove difficult, they attributed this lack of difficulty to a tabular, "nearly wordless" approach. Shneiderman (1978) discusses the difficulty associated with universal quantification. He also compared artificial and natural languages for querying, finding that the latter led to more questions which could not be answered from the database. In another echo of the problems experienced with ontologies, Jagadish et al. (2007) commented that database users were having difficulty with "the mere complexity of the … schema". They also commented on the differing

expectations of database use and web search. For example, database systems need to be able to explain their results; a feature analogous to providing justifications in ontology debugging. However, to the authors' knowledge, there has been little exploitation in database research of ideas from theories of reasoning or philosophy of language.

Whilst databases are a long-established area of research, the Linked Open Data (LOD) cloud is a new one. It is characterised by relatively simple schema, often using only the constructs available in RDFS. The interrogation mechanism is frequently the SPARQL query language. There is evidence that the great majority of queries are relatively simple (Gallego et al. 2011; Möller et al. 2010). It is an open question whether this is because such queries are all that users need; all that they can conceive of; or all that they can easily construct using SPARQL. As with databases and DL, there have been experiments with the use of natural language (Chang et al. 2015; Rico et al. 2015). Valid research questions are how natural language and visual query languages compare with SPARQL, and to what extent one can apply the insights from psychology and language discussed in this paper.

8.5 Final comments

This paper began by noting the adoption of DLs as the dominant KR languages, in place of the more psychologically-inspired frame-based paradigm. Subsequently, there has been very considerable attention to the logical and computational properties of DLs, but very little attention to their usability. The major attempt to address usability has been through the adoption of MOS. Yet Alharbi et al. (2017) have called into question the effectiveness of MOS when compared with the German DL syntax. Moreover, the work reported here has illustrated that some MOS keywords may be ambiguous. Additionally, the work of Sarker et al. (2017) has challenged the structure of all DL syntaxes, proposing a rule syntax for DL.

This paper has shown that insights from cognitive psychology and language studies can help to understand the difficulties experienced with knowledge representation languages, and lead to suggestions to mitigate those difficulties. In the future it is hoped that this approach will be extended, e.g. to query languages, and will draw on more recent theories of reasoning, e.g. probabilistic and non-monotonic.

Acknowledgements

The authors wish to thank all the study participants, and particularly Dr Gem Stapleton and Dr John Davies for organising study sessions at their respective institutions. The authors would also like to thank the anonymous reviewers for various comments which have been incorporated into the paper.

References

Alharbi, E., Howse, J., Stapleton, G., Hamie, A., & Touloumis, A. (2017). The Efficacy of OWL and DL on User Understanding of Axioms and Their Entailments. In *International Semantic Web Conference* (pp. 20–36). Springer.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An Introduction to Description Logic*. Cambridge University Press.

Baader, F., & Nutt, W. (2003). Basic description logics. In *Description logic handbook* (pp. 43–95).

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1.

Braine, M. D. S., & O'Brien, D. P. (1998). How to investigate mental logic and the syntax of thought. In *Mental Logic* (pp. 45–61). Mahwah, New Jersey: Lawrence Erlbaum.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83.

Chang, B. K. W., Lefevre, M., Guin, N., & Champin, P.-A. (2015). SPARE-LNC: un langage naturel contrôlé pour l'interrogation de traces d'interactions stockées dans une base RDF. In *IC2015*. Rennes, France.

Craik, K. J. W. (1967). *The nature of explanation*. CUP Archive.

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 296–306.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, *54*(1), 1–71.

Gallego, M. A., Fernández, J. D., Martínez-Prieto, M. A., & de la Fuente, P. (2011). An empirical study of real-world SPARQL queries. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hyderabad, India*.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan, *Syntax and Semantics, Volume 3: Speech Acts* (pp. 41–58).

Halford, G. S, Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(06), 803–831.

Halford, Graeme S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*(1), 70–76.

Hopkins, W., Marshall, S., Batterham, A., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine+ Science in Sports+ Exercise*, *41*(1), 3.

Horridge, M., Bail, S., Parsia, B., & Sattler, U. (2011). The cognitive complexity of OWL justifications. *The Semantic Web–ISWC 2011*, 241–256.

Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H. (2006). The manchester owl syntax. *OWL: Experiences and Directions*.

Horridge, Matthew, & Patel-Schneider, P. F. (2008). Manchester syntax for OWL 1.1. *OWL: Experiences and Directions, Washington, DC*.

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., & Yu, C. (2007). Making database systems usable. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 13–24). ACM.

Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (p. 179). Psychology Press.

Johnson-Laird, P. N. (2010). Against logical form. *Psychologica Belgica*, *50*(3), 193–221.

Johnson-Laird, Philip N. (2005). Mental models and thought. *The Cambridge Handbook of Thinking and Reasoning*, 185–208.

Johnson-Laird, Philip N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418.

Johnson-Laird, Philip N., Byrne, R. M., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review*, *96*(4), 658.

Johnson-Laird, Philip Nicholas, & Byrne, R. M. (1991). *Deduction.* Lawrence Erlbaum Associates, Inc.

Khan, M. T., & Blomqvist, E. (2010). Ontology design pattern detection-initial method and usage scenarios. In *SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing* (pp. 19–24).

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). *Negating compound sentences*. Naval Research Lab, Washington DC, Navy Center for Applied Research in Artificial Intelligence. Retrieved from http://mindmodeling.org/cogsci2012/papers/0110/paper0110.pdf

Kifer, M., Lausen, G., & Wu, J. (1995). Logical foundations of object-oriented and frame-based languages. *Journal of the ACM (JACM)*, *42*(4), 741–843.

Krötzsch, M., Simancik, F., & Horrocks, I. (2012). A description logic primer. *ArXiv Preprint ArXiv:1201.4089*.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, *40*(1), 121–170.

Lassila, O., & McGuinness, D. (2001). *The Role of Frame-Based Representation on the Semantic Web*. Stanford University. Retrieved from http://www-ksl.stanford.edu/pub/KSL_Reports/KSL-01-02.html

Mendonça, E. A., Cimino, J. J., Campbell, K. E., & Spackman, K. A. (1998). Reproducibility of interpreting" and" and" or" in terminology systems. In *Proceedings of the AMIA Symposium* (p. 790). American Medical Informatics Association.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211–277). McGraw-Hill.

Möller, K., Hausenblas, M., Cyganiak, R., Grimnes, G., & Handschuh, S. (2010). Learning from Linked Open Data Usage: Patterns & Metrics. In *WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC, US.

Nguyen, Power, Piwek, & Williams. (2012). Measuring the understandability of deduction rules for OWL. Presented at the First international workshop on debugging ontologies and ontology mappings, Galway, Ireland.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Partee, B., & Rooth, M. (1983). Generalized conjunction and type ambiguity. *Formal Semantics: The Essential Readings*, 334–356.

Power, R. (2010). Complexity assumptions in ontology verbalisation. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 132–136).

Power, R., & Third, A. (2010). Expressing OWL axioms by English sentences: dubious in theory, feasible in practice. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1006–1013).

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. ISBN 3-900051-07-0.

Ragni, M., Eichhorn, C., & Kern-Isberner, G. (2016). Simulating Human Inferences in the Light of New Information: A Formal Analysis. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI*.

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., … Wroe, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web* (pp. 63–81). Springer.

Rector, A. L. (2003). Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. In *Pacific Symposium on Biocomputing* (pp. 226–237).

Rico, M., Unger, C., & Cimiano, P. (2015). Sorry, I only speak natural language: a pattern-based, data-driven and guided approach to mapping natural language to SPARQL. In *Intelligent Exploration of Semantic Data (IESD) 2015*.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*(1), 38.

Sarker, M. K., Krisnadhi, A., Carral, D., & Hitzler, P. (2017). Rule-Based OWL Modeling with ROWLTab Protégé Plugin. In *European Semantic Web Conference* (pp. 419–433). Springer.

Schwitter, R., Kaljurand, K., Cregan, A., Dolbear, C., & Hart, G. (2008). A comparison of three controlled natural languages for OWL 1.1. In *4th OWL Experiences and Directions Workshop (OWLED 2008 DC), Washington*.

Scott, D. (2012). Tukey's ladder of powers. *Rice University*. Retrieved from http://onlinestatbook.com/2/transformations/tukey.html

Shneiderman, B. (1978). Improving the human factors aspect of database interactions. *ACM Transactions on Database Systems (TODS)*, *3*(4), 417–439.

Stapleton, G., Howse, J., Bonnington, A., & Burton, J. (2014). A vision for diagrammatic ontology engineering. Retrieved from http://eprints.brighton.ac.uk/13046/

Stapleton, G., Howse, J., Taylor, K., Delaney, A., Burton, J., & Chapman, P. (2013). Towards Diagrammatic Ontology Patterns. Presented at the 4th Workshop on Ontology and Semantic Web Patterns, Sydney, Australia. Retrieved from http://ontologydesignpatterns.org/wiki/WOP:2013

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*, 109–159.

Stevens, R., Aranguren, M. E., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., & Rector, A. (2007). Using OWL to model biological knowledge. *International Journal of Human-Computer Studies*, *65*(7), 583–594.

Thomas, J. C., & Gould, J. D. (1975). A Psychological Study of Query by Example. In *Proceedings of the May 19-22, 1975, National Computer Conference and Exposition* (pp. 439–445). New York, NY, USA: ACM. https://doi.org/10.1145/1499949.1500035

van Schaik, P., & Weston, M. (2016). Magnitude-based inference and its application in user research. *International Journal of Human-Computer Studies*, *88*, 38–50.

Vigo, M., Jay, C., & Stevens, R. (2014). Protege4US: harvesting ontology authoring data with Protege. Presented at the HSWI2014 - Human Semantic Web Interaction Workshop, Crete.

W3C. (2001). Web Ontology Language (OWL). Retrieved from http://www.w3.org/2001/sw/wiki/OWL

Warren, P. (2017). *Human reasoning and Description Logics - applying psychological theory to understand and improve the usability of Description Logics* (Ph.D. dissertation). Open University (U.K.).

Warren, P., Mulholland, P, Collins, T, & Motta, E. (2014). The usability of Description Logics: understanding the cognitive difficulties presented by Description Logics (pp. 550–564). Presented at the ESWC 2014, Crete: Springer.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2014). Using ontologies. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 579–590). Springer.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2015). Making sense of description logics. In *Proceedings of the 11th International Conference on Semantic Systems* (pp. 49–56). ACM.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2017). Improving the Comprehensibility of Description Logics - Applying insights from theories of reasoning and language. Presented at the ESWC 2017, Portoroz, Slovenia: Springer.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul.